te technical note techn

The Development and Evaluation of a Behaviorally Based Rating Form for Assessing Air Traffic Controller Performance

Randy L. Sollenberger, ACT-530
Earl S. Stein, ACT-530
and
Stan Gromelski, PERI

February 1997

DOT/FAA/CT-TN96/16

U.S. Department of Transportation
Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/CT-TN96/16		

4. Title and Subtitle	5. Report Date
The Development and Evaluation of a Behaviorally Based Rating Form for Assessing Air Traffic Controller Performance	February 1997
	6. Performing Organization Code ACT-530
7. Author(s) Randy Sollenberger and Earl Stein, ACT-530, and Stan Gromelski, PERI	8. Performing Organization Report No. DOT/FAA/CT-TN96/16
9. Performing Organization Name and Address Federal Aviation Administration William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	10. Work Unit No. (TRAIS) 11. Contract or Grant No. F22022
12. Sponsoring Agency Name and Address Federal Aviation Administration Human Factors Division 800 Independence Avenue, S.W. Washington, DC 20591	13. Type of Report and Period Covered Technical Note 14. Sponsoring Agency Code AAR-100
15. Supplementary Notes	

16. Abstract

The evaluation of air traffic controller performance is a complex process. While there are standard forms in field use, there is currently no comprehensive system for reliable observer evaluation. This research involves the development of a new form along with a training package for use in research and possibly operational testing. The form consists of 24 rating scales. These scales focus on observable actions that trained air traffic control specialists could identify to make behaviorally based ratings. The study evaluates the reliability of the rating form by determining the consistency of ratings obtained from six observers who viewed videotapes of a previously recorded simulation study. These observers were supervisors and training staff specialists from Terminal Radar Approach Control facilities nationwide. Prior to making formal ratings, the observers participated in a training program designed to help them become proficient in observational rating. During the evaluation phase of the study, the observers viewed 20 one-hour videotapes of controllers working different traffic scenarios. The results indicated that most of the rating scales had reasonable inter-rater reliabilities ranging from $r = .7$ to $r = .9$. The study also identified the performance areas that were more difficult for observers to evaluate consistently, possibly due to misunderstanding rating criteria or overlooking critical controller actions.

17. Key Words	18. Distribution Statement
Controller performance measurement Performance rating form Measurement reliability Videotape evaluation	This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	61	

Table of Contents

List of Illustrations

Figures Page

Tables Page

Executive Summary

The purpose of this research was to develop and assess a performance evaluation method intended to provide a comprehensive assessment of air traffic controller performance. A rating form was designed to be used as a testing and evaluation tool to measure the effectiveness of new Air Traffic Control (ATC) systems, system enhancements, and operational procedures in simulation research. The focus of the rating form was on observable actions that trained air traffic control specialists (ATCSs) could identify to make behaviorally based ratings of controller performance. The present study evaluated the reliability of the rating form by determining the consistency of ratings obtained from six observers who viewed videotapes of controllers from a previously recorded simulation study.

The rating form used in the present study was based on a form recently developed by Hedge, Borman, Hanson, Carter and Nelson (1993) working on a Federal Aviation Administration (FAA) project titled, "Separation and Control Hiring Assessment (SACHA)." Although the SACHA project goals are different, the research that formed the basis of their rating scales was very useful in developing the present rating form.

The rating form was developed through preliminary work with seven ATCSs, who either reviewed drafts of the rating form or actually used the form to evaluate controllers. The rating form contained 24 rating scales assessing different areas of controller performance. The performance areas were organized into six performance categories, with an overall rating scale for each category. Several controller actions were identified for each of the performance areas. These controller actions were observable behaviors that ATCSs should always look for when evaluating controllers.

The rating form was constructed with an 8-point rating scale format, with statements describing the necessary controller actions for each scale point. The ATCSs recognized the importance of the scale point descriptions and spent much time working to improve the terminology. A comment section was included for each of the performance areas. The comment sections were used for describing the effective and ineffective controller actions that were observed. The comments served as a justification for the ratings that were given and helped the research team understand the observations that led to each rating.

The videotapes used in the present study were recorded during a simulation study conducted by Guttman, Stein, and Gromelski (1995) with Atlantic City International Airport Terminal Radar Approach Control (TRACON) controllers. The purpose of the Guttman et al. study was to develop and validate a generic TRACON that would be used as a standard testing environment in future ATC simulations.

The study was conducted at the FAA Research Development and Human Factors Laboratory at the William J. Hughes Technical Center at the Atlantic City International Airport in New Jersey. Six TRACON supervisors and training staff specialists from different ATC facilities nationwide participated as observers. The participants watched two views of the previously recorded simulations. The first view was an over-the-shoulder recording of the controller's upper body

that showed interactions with the workstation equipment. The second view was a graphical playback of the traffic scenario that showed all the information on the controller's radar display. Both views were simultaneously presented on different screens and synchronized with an audio recording of communications between the controllers and simulation pilots.

The observers participated in a training program before using the rating form to make formal evaluations of the videotapes. The training program was designed to help the observers learn the airspace in the simulation and become proficient with the rating form. Several steps encouraged the observers to adopt mutual evaluation criteria for their ratings. First, the research team discussed common rater biases and how to avoid them. Then, the observers reviewed the rating form and discussed their interpretations of its terminology. Next, the observers used the rating form while viewing six practice videotapes. After each tape was viewed, the observers discussed what they saw and why they selected their ratings. The discussions helped to clarify some of the ambiguities in the rating form and identify the observers whose rating behavior differed a great deal from the others.

The observers completed the training program in one week. The actual videotape evaluations were completed the second week. The researchers randomly selected 4 of the 10 controllers who participated in the generic sector study and used all 4 videotapes from each controller. Additionally, the observers viewed one tape from each of the controllers a second time to obtain a measure of reliability on repeated occasions. In total, the observers viewed 20 one-hour videotapes, with 5 tapes shown on each day. On the last day of the study, the observers completed a questionnaire that asked them to provide weighting values indicating the relative importance of the six performance categories. The weights were used to calculate an overall performance score for the controller in each of the videotapes. Finally, observers answered a few questions about the training program and methods used in the study.

The researchers assessed two types of reliability: inter-rater reliability and intra-rater reliability. Inter-rater reliability refers to the consistency of the ratings between the observers, and intra-rater reliability refers to the consistency of the ratings on repeated occasions. The results indicated that most of the rating scales were very reliable, although there were a few exceptions. In addition to the subjective ratings obtained from the observers, the present study examined several system effectiveness measures (SEMs) that are routinely collected in ATC simulation research. The SEMs included the number of conflict errors, controller assignments, controller transmissions, aircraft density, total aircraft distance flown, and controller workload. The study identified the performance areas that were more difficult for observers to evaluate consistently, possibly due to misunderstanding the rating criteria or overlooking critical controller actions. The study identified the number of ground-to-air transmissions as the SEM most strongly related to overall controller performance. Additionally, several individual performance areas were strongly related to overall controller performance. The study demonstrated the feasibility of using videotapes as a presentation method for evaluating controller performance. Finally, observers accepted the new rating form and suggested only a few changes to improve the organization and terminology.

1. Introduction

1.1 Problem Statement

Human performance in a complex system is an essential component of overall system performance. When human beings are in the command and control loop, the decisions they make and how well they carry them out have a direct impact on the degree that the system can achieve its goals. However, there is often some disagreement on what role the human really plays in a system and what constitutes performance. Some systems are more error tolerant and forgiving than others. Most systems do have some definition of minimum necessary performance for their operators, but they do not differentiate well when it comes to various levels of performance quality above the minimum level. In Air Traffic Control (ATC), minimum standards are mandated by safety considerations, laws, and Federal Aviation Regulations (FARs). Beyond that, there is considerable variance in opinions about what constitutes good, better, and best human performance.

The Federal Aviation Administration (FAA) William J. Hughes Technical Center has been examining performance issues for over 30 years and is a leader in ATC simulation. Much of the measurement capability that has evolved over the years emphasizes system effectiveness measures (SEMs) that can be collected in real time during ATC simulations. SEMs are objective measures that can be collected and analyzed to evaluate the effects of new systems and procedures. However, objective measures cannot encompass the full range of factors that describe good and poor ATC and may fail to capture the essence of controller performance.

1.2 Assumptions and Goals

This study was developed to determine if the objective measures that can be collected are related to how a panel of subject matter experts (SMEs) view controller performance. No measurement system in applied science is valuable if it does not make some sense to the people who have to use it to make decisions. The problem for this study was twofold. First, could a group of SMEs be trained to evaluate air traffic controller performance so that they were all looking for the same types of behaviors and placing similar values on them? The second issue was only relevant if the first issue proved to be true. If SMEs could agree on their evaluation of performance, would their pooled evaluations be related to the objective performance data collected in a simulation environment?

This research study began with the belief that it is possible to train supervisory air traffic controllers to objectively observe and evaluate behavior. They all have experience in using FAA Form 3120-25, the ATCT/ARTCC OJT Instructor Evaluation Report. This research assumed that this form could be improved and, when supported by a training program, the quality of the ratings will also improve. The primary criteria of rating quality are various measures of inter-rater reliability. This research was not intended to replace FAA Form 3120-25. Rather, its purpose was to develop an observational rating system with associated rater training that could be used later to validate other measurement systems.

There are many definitions of performance. Bailey (1982, p.4) defined performance as follows: "Performance then is defined as the result of a pattern of actions carried out to satisfy an objective according to some standard. These actions may include observable behavior or non-observable intellectual processing (e.g., problem solving, decision making, planning, reasoning). Things change when people perform." For the purposes of this research, the operational definition of performance is the accomplishment of a task or interrelated set of tasks in relation to a defined and specified standard while operating within constraints of space, time, and resources. The concept of performance implies that it can vary along a continuum of quality based on a wide variety of variables. One critical variable is the human operator.

The operator must accomplish specified tasks that are evaluated in relationship to a specified standard. If the behavior exceeds the standard, it is evaluated as successful. If the behavior fails to meet the standard, it is not successful. The distance above or below the standard determines different levels of accomplishment within the successful and unsuccessful categories, respectively. The evaluation is more difficult when it must take into account the relative levels above or below the absolute standard.

This research is based on the belief that either there is an absolute standard for ATC performance or that experts can agree on a more relative subjective standard. It also assumes that these experts can apply the relative standard in some consistent way.

1.3 Review of Related Literature

Researchers and personnel specialists find it difficult to develop performance criteria. Thorndike (1982) suggested that criteria should define success on the job, and that is one area where many problems occur. Thorndike stated, "It is difficult even to formulate any complete definition of success on the job, much less develop a measure that adequately represents it" (p. 14). Most performance indicators are partial and incomplete. According to Thorndike, they lack range and time span. They only provide a snapshot. Criteria can be confounded by irrelevant sources of variance, such as rater biases and low or unknown reliability. There are relatively few jobs where a performance test is appropriate. It is necessary to determine what behaviors best represent the skill or what aspects of a product should be evaluated to determine performance. Thorndike concluded that "performance evaluation (in many settings) tends to be subjective and unreliable at best" (p. 27).

ATC involves both individual and team performance to keep the system functioning smoothly. Performance criteria have been and continue to be a challenge. In 1994, there were 772 air traffic controller operational errors, which seems to be a large number (FAA, 1995). This translates into a rate of 0.53 errors per 100,000 facility operations, which is actually a very small percentage. The measures that are employed often convey a different meaning concerning the quality of system or individual performance. Also, errors vary considerably in terms of their severity. A simple tally of errors by type does not truly convey what is going on in the system. More information is needed about the nature of these errors.

In an early, comprehensive study of controller errors, Kinney, Spahn, and Amato (1977) analyzed FAA reports and developed eight categories of errors. These include (a) controlling in

another controller's airspace, (b) timing and completeness of flight data handling, (c) interpositional coordination of data, (d) use of altitude on the display, (e) procedures for scanning and observing flight data, (f) phraseology and use of voice communications, (g) use of human memory to include relying on recall in a noisy environment, and (h) dependence on automatic capabilities. The work of Kinney and his associates was based on a considerable amount of data collected in operational environments. The taxonomy has had an impact on error evaluation in research, but did not become the FAA standard for classifying operational errors.

Today, the FAA uses a different set of categories to do this classification. The following categories were employed by the FAA (1988): radar display, communication, coordination, aircraft observation, data posting, and position relief. By far, the most frequent source of errors was in the subclass of radar display: the misuse of data. This category implies that information was available and was either misinterpreted or inaccurately stored in working memory.

Rodgers (1993) analyzed the FAA operational error data base and found that facility error rates were inversely proportional to the percentage of the work force that had achieved full performance level (FPL) status. However, in terms of evaluating new systems or personnel who have already achieved FPL status, operational errors are an imprecise metric and are not very useful when applied by themselves. Controllers still demonstrate varied performance despite meeting minimum standards (by not committing errors most of the time). Since errors alone have not been very effective data sources, scientists have attempted to develop more complex multivariate performance models. These have sometimes taken the form of fast time computer simulation models.

Robertson, Grossberg, and Richards (1979) developed and evaluated such a computer model of controller activity. This model became known as the relative capacity estimating process (RECEP). It included both workload and performance variables. The model emphasized system events and functions in an off-line data processor capable of analyzing these events after they have occurred. While the primary purpose of the model was to estimate workload, it examined three general categories of controller activities: routine, surveillance, and conflict prevention. By computing and summing all the subtask performance times, the authors proposed maximum limits for man minutes-per-hour of operational time. RECEP measures correlated favorably with SMEs' ratings of workpace. There was considerable RECEP variability across different airspace sectors indicating that airspace structure may be one of the influencing factors in controller performance. Another factor could involve how controllers approach the environment and analyze the situation.

As part of a larger project aimed at improving controller training, a group of researchers performed a cognitive task analysis of expertise to see if experts and novices differed in how they think (Seamster, Redding, Cannon, Ryder, & Purcell, 1993). They concluded that experts took a wider view of the evolving air traffic situation. Experts appear to be more flexible in their approach to the dynamics in their airspace. The researchers identified 13 en route controller tasks that were linked to their cognitive models of the airspace. These were (a) maintain situation awareness, (b) develop and revise sector control plan, (c) resolve aircraft conflicts, (d) reroute aircraft, (e) manage arrivals, (f) manage departures, (g) manage overflights, (h) receive handoffs, (i) receive pointouts, (j) initiate handoffs, (k) initiate pointouts, (l) issue advisories, and

(m) issue safety alerts. Each of these tasks is broken into numerous subgoals that establish the matrix of the controller's mental model.

According to Seamster et al. (1993), their research supports the hypothesis that experienced controllers group or organize their "picture" by events rather than by individual aircraft. The mental model and task accomplishment or requirement interact and influence each other. When thinking out complex ATC problems, experts used fewer but more detailed planning strategies while maintaining more alternatives for managing workload than those available to less experienced controllers.

Endsley and Rodgers (1994) studied en route ATC from the viewpoint of the requirements generated for situation awareness, another cognitive approach. These researchers attempted to identify the essential components of information that an en route controller must have in situation awareness to perform their tasks. They chose to work backward from major operational goals through subgoals. This was a cognitive study rather than a task analysis. Using a panel of eight SMEs, the researchers employed a replay of ATC incidents to cue participant memory. The end product of this work was a series of information requirements linked to each aspect of the controller's duties. This may have implications for future performance evaluation if the presence or absence of these elements of information is reflected in actual performance. This is an example of using data bases that are available to produce models and concepts. Generating new data under controlled conditions can also be useful for understanding controller performance.

Simulation research has been used to study ATC equipment, procedures, and concepts for over 35 years. Over this period, various sets of dependent variables have evolved to assist in the evaluation of system and individual controller performance. The specific subset of variables has generally been tailored to meet the research goals of each study. Most of the ATC simulation studies have been conducted at the FAA William J. Hughes Technical Center. Buckley, O'Connor, Beebe, Adams, and MacDonald (1969) conducted a simulation study focused on the assessment of controller performance and its relationship to chronological age. Buckley and his colleagues were among the few researchers who have used a combination of objective system measures and over-the-shoulder SME ratings. They commented that a difficulty with subjective ratings is their frequent unreliability. They employed eight observers who did over-the-shoulder ratings in pairs. These observers were current controllers from facilities other than those where the participants worked. Using intraclass correlations as the indicator of inter-rater reliability, the correlations between pairs of raters ranged from .06 to .72.

Buckley, DeBaryshe, Hitchner, and Kohn (1983) performed two experiments to examine the use of simulation for evaluating air traffic controller performance. They emphasized the quality of measurement and identified the basic dimensions for measuring ATC functions in real time. This first experiment examined the effects of using two en route sector layouts and three traffic density levels ranging from very light to very heavy. Data were collected from two 1-hour runs for each of 31 controllers. In the first experiment, there were statistically significant effects of sector geometry and traffic density for almost all the 10 performance measures. There was also a significant interaction effect between geometry and density. Sector geometry appeared to have a major impact on controller performance. This led to the design of a second experiment.

The second experiment examined the effects of collecting data over time by repeated measures. Twelve 1-hour runs were conducted using the same sector with the same traffic level for each of 39 controllers. A factor analysis was computed to look for redundancy in the measures used to quantify system performance. This produced four meaningful factors or measures: confliction, occupancy, communication, and delay. The confliction factor included measures of 3-, 4-, and 5-mile conflicts. The occupancy factor included measures of the time an aircraft was under control, distance flown under control, fuel consumption under control, and time within boundary. The communications factor included path changes, number of ground-to-air communications, and the duration of ground-to-air communications. The delay factor included total number of delays (aircraft delayed en route by controller actions) and total delay times. Two auxiliary measures, number of aircraft handled and fuel consumption, were also relevant. The data resulting from the first experiment of Buckley et al. (1983) were cross-validated with the factor analysis derived from the second experiment. These experiments conducted by Buckley et al. have served as building blocks for most of the controller performance research that followed using both simulation and field facility research.

Researchers in ATC performance have typically developed their own measurement tools that were tailored to their immediate and long-term needs. Stein and Buckley (1992) assembled and consolidated the variables that had been useful over the years for researchers at the FAA William J. Hughes Technical Center. This work was based primarily on the research of Buckley et al. (1983) and, to a lesser extent, on research accomplished by Stein (1984a, 1984b, 1985). The majority of the performance measures are based on frequencies of events and time, both of which may be summed over any specified period. These frequency performance measures have been used in numerous studies over the years to evaluate concepts and systems. However, researchers cannot always clearly define the difference between system and individual performance measures. The two are usually integrated in complex ways within any given study.

In one study, researchers compared parallel approach separation standards between 1.5 and 2.0 nmi. The variables measured included controller operational errors and landing rates at the airport under study. The results demonstrated that controller performance in terms of error frequency and landing rates did not decline and there was no increase in subjective self-reported estimates of workload. The landing rates, possibly a system variable, were higher for the reduced separation standard (Stein, 1989).

In a more recent study, Sollenberger and Stein (1995) conducted a study of controller memory issues to determine whether performance could be enhanced using a memory aid. The performance measures were collected automatically when each of 16 controllers worked in simulated Terminal Radar Approach Control (TRACON) airspace. The memory aids had some positive influence on controller's behavior. In the aided condition, controllers made significantly fewer ground-to-air transmissions and gave fewer altitude and heading changes. Communication variables like these have been used as indicators of controller workload in other studies (Robertson et al., 1979). Another positive result was that under the memory-aided condition controllers made fewer hand-off errors.

Guttman, Stein, and Gromelski (1995) recently completed a performance-based study. Controllers worked under two sets of airspace conditions, one with which they were familiar and

one that was designed to be a generic terminal radar approach model. The purpose of this study was to evaluate controller performance under both conditions and to see if the generic model could be used for future research and training purposes. A wide variety of objective and subjective data was collected. Controllers were able to learn the generic airspace rather quickly, and performance variables did not change appreciably over the course of familiarization with the generic sector. The generic sector was easy to learn and did not lead to performance decrements. Over-the-shoulder observers also rated the performance of the participants and estimated how hard they were working. These observations were consistent with the objective data in that they showed that there were few differences in performance between home and generic sectors. The majority of the participating controllers indicated that the airspaces were both realistic and representative of the TRACON environment.

1.4 Observing and Rating Behavior

Controller performance measurements have consistently involved tasks and variables derived from ATC and produced findings expressed in ATC terms (Hopkin, 1980). Hopkin believed that it was also important to use basic psychological knowledge to explain controller behavior. He also believed the controller's task should be considered in human terms to provide perspectives, explanations, and insights into the cognitive processes that support ATC. Hopkin (1991) indicated that, in the long run, we may have to expand the more traditional views of what performance is to encompass concepts that we have previously ignored as unrelated or inconsequential.

Subject matter expertise and knowledge are basic requirements for evaluating the performance of others. However, sometimes SMEs are tempted to apply a personal standard, or "my standard," rather than the designated standard. My standard is influenced by the SME's experience, training, performance of current peers, and possibly by the organizational standards (Anastasi, 1988).

Much of the literature on performance evaluation is based on performance appraisals, which are accomplished in organizations on an annual or semi-annual basis. These are heavily dependent on the rater's memory for events. There are common rating errors that reduce reliability and validity. These include but are not limited to halo effects, leniency, stringency, central tendency errors, and primacy effects (Bass & Barrett, 1981). Even in memory-dependent, organizational-type performance appraisals, training raters can reduce the effects of leniency and halo effects (Anastasi, 1988).

Performance rating in real time is less dependent on memory. Real-time ratings suffer from all the biases cited previously but can be more focused on actual behavior. Performance appraisals in business and industry may be accomplished for very different reasons than ratings used for human factors purposes. In many organizations, such appraisals are used for compensation, promotion, and retention purposes (Bender, Eichel, & Bender, 1985). While, in theory, the purpose of the rating should not influence the quality of the rating design or implementation, in practice, it might. Questions concerning reliability and validity are less likely to be raised than in a human factors evaluation. Training is an area where there may be an overlap. Training results could have direct and immediate impact on organizational performance. However, ratings are

often done unsystematically, without adequate scale development or rater preparation. Organizations often opt for a simplistic approach to performance, identifying it for the presence or absence of error. A controversy continues to exist over the relative merits of observational rating as compared to more objective data that could be collected in a laboratory.

Hennessy (1990) identified two major approaches to human performance measurement that should be discouraged: trying to measure humans like machines and attempting to move the laboratory to the field environment. He also noted that good performance measurement for real-world environments does not yet exist. The reasons for these issues are that researchers overemphasize the appearance of objectivity and the automaticity of measurement. Objectivity is often equated with the ability to collect data with machines. For example, computerized measurement of pilot proficiency has not proved to be useful as ratings by instructor pilots. Hennessy believes that the future should involve more observational rating and less laboratory assessment.

Anastasi (1988) discussed the use of ratings as criterion measures for the validation of other primarily predictive indicators. She commented that despite the technical shortcomings and the biases of observers, ratings can be valuable sources of criterion information when they are collected under systematic conditions. Anastasi stressed the importance of observer/rater training to increase reliability and validity while reducing common judgmental errors. This training can take many forms, but anything that enhances a rater's observational skills will most likely improve the quality of the ratings.

Controllers have used over-the-shoulder ratings since the beginning of the ATC system. They express the belief in their ability to observe and evaluate each other. Careers may be influenced, especially during training, based on the ratings and comments placed on FAA Form 3120-25. The form contains 27 scales, divided into 5 categories: Separation, Control Judgment, Methods and Procedures, Equipment, and Communication/Coordination. Each scale allows for a rating on three points: Satisfactory, Needs Improvement, and Unsatisfactory. There is no space for written observations on the front side of the form. Written observations are reserved for the back. Controller culture is such that, when receiving a "check ride" by a trainer or supervisor, if they notice him/her writing, they become worried (G. Bing, personal communication, February 15, 1996). Writing is discouraged unless something is wrong. The rater likely experiences some subtle pressure to avoid writing and to depend on his/her memory for events related to performance. Depending on memory and using a 3-point scale are basic prescriptions for unreliable measurement.

Controllers tend to be very decisive individuals, and it can be difficult to change their assumptions about rating performance. When observing the same behavior, at the same time under the same conditions, well-meaning observers who have not been trained to systematically observe may generate very different results. Under such circumstances, inter-rater reliability can break down. This actually was to happen in this study during the beginning of training.

Observer ratings have frequently been used in ATC simulation research. Boone and Steen (1981) compared computer-derived measurements with more traditional over-the-shoulder methods using ATC students. They employed five observer/instructors and 48 student

controllers. There was an emphasis in the observing/rating process of identifying student performance errors. The inter-rater reliability for the observers were relatively low, ranging from .23 to .58. Regression analysis of the computer scores against a global rating of student potential led to a multiple *R* of .52. The authors speculated that the computer-generated scores could potentially be used to predict on-the-job success.

In their comprehensive study of SEMs, Buckley et al. (1983) included ratings as part of the overall measurement package. Two observers were asked to complete ratings every 10 minutes during the simulations. They used a 10-point scale to rate two areas: overall system effectiveness and individual controller judgment/technique. Buckley et al. evaluated inter-rater reliability using intraclass correlations, which ranged from .06 to .72. While not cited directly in text, the median inter-rater reliability appeared to be around .60. Individual correlations between observer ratings and the SEMs spread around a median of .25. Multiple regressions of observer ratings and major SEM factors produced multiple *R*s that ranged around $r = .70$.

Stein (1984c) conducted a real-time simulation with 10 air traffic controllers who worked under three levels of taskload. Along with collecting automated performance measures, which included some of the same ones that Boone and Steen (1981) had employed, two trained observers independently evaluated participant's workload, busyness, and effectiveness. Inter-rater reliability was very high ($r = .91$), and observed workload was strongly related to task load as defined by variables such as average instantaneous aircraft count. There was an inverse relationship between ratings on workload and effectiveness ($r = -.55$). This type of inverse relationship is not uncommon under conditions where controllers work traffic across a wide range of task loads.

Ratings are often used as an additional source of data in ATC simulations. For example, Sollenberger and Stein (1995) employed an SME to observe and rate workload and performance of controllers during a simulation study testing memory aids. Only one expert was available, so no estimate of inter-rater reliability was possible. The observer's ratings of workload correlated with the controllers real-time workload ratings ($r = .85$). The observer's workload ratings were inversely related to performance ratings ($r = -.54$) and to airspace complexity measures ($r = -.56$). These correlations were significant from zero and were similar to findings from other studies accomplished at the FAA William J. Hughes Technical Center (see Stein, 1985).

2. Experiment

2.1 Purpose

The purpose of this research was to develop and evaluate a performance evaluation method intended to provide a comprehensive assessment of air traffic controller performance. The rating form was designed as a research tool to measure the effectiveness of new ATC systems, system enhancements, and operational procedures in simulation research. The rating form was not designed for technical performance appraisals of controllers at field facilities or to select ATC trainees for the academy, but could potentially be used for these purposes with modifications. The focus of the rating form was on observable actions that trained air traffic control specialists (ATCSs) could identify to make behaviorally based ratings of controller performance. The

present study evaluated the reliability of the rating form by determining the consistency of ratings obtained from six observers who viewed videotapes of controllers from a previously recorded simulation study.

2.2 Rating Form Development

The rating form in the present study was based on another form recently developed by Hedge, Borman, Hanson, Carter and Nelson (1993), working on the Separation and Control Hiring Assessment (SACHA) project. One of the SACHA project tasks was to develop a set of rating scales based upon the job requirements of controllers and to use the scales as a measurement system for assessing controller performance. The SACHA team had SMEs generate specific examples of effective and ineffective controller performance. The performance examples were then grouped into 10 performance categories. Each example was rated for effectiveness on a scale from 1 (very ineffective) to 7 (very effective). Based upon the performance examples and ratings, summary statements were generated that described ineffective, average, and highly effective performance in each of the performance categories. The final version of the present rating form is shown in Appendix A, and a draft of the SACHA rating form is shown in Appendix B.

Several design modifications were made to the SACHA rating form to meet research requirements. The rating form in the present study was developed through preliminary work with seven ATCSs who either reviewed drafts of the rating form or used the form to evaluate controllers in video-recorded simulations. Although the ATCSs agreed that the 10 performance categories adequately covered the major aspects of ATC, 3 of the categories were omitted from the final version of the rating form. Specifically, coordinating and teamwork were omitted because controller actions in these categories were either simplified or not present in the simulations. A third category, reacting to stress, was omitted because performance in this category could not be reliably observed in the actions of the experienced controllers who participated in the simulations.

After some preliminary work, the researchers decided that increasing the number of rating scales would improve the rating form. The performance categories remained as an effective method for organizing the rating scales, but each category was divided into different performance areas. Based upon the SACHA performance examples, it seemed reasonable to construct specific performance areas that were related to the general category but sufficiently different from each other to be included as separate rating scales. The final version of the form contained 24 rating scales assessing different areas of controller performance. This modification avoided the "mixing apples and oranges" problem, as one ATCS called it, of making a single rating about controller actions completed with different levels of effectiveness. A large number of rating scales is desirable for research purposes to identify the specific performance areas that are affected by a proposed change to the ATC system. Also, the researchers designed an overall rating scale for each performance category to have generality and specificity in the ratings.

Measurement sensitivity, which is the ability to detect small differences in performance, is another desirable feature for research purposes. Sensitivity is important because a proposed change that improves controller performance even slightly in simulations may have a major

impact on the ATC system in the long term. Increasing the number of rating scale points is one potential technique for improving measurement sensitivity. However, this technique will increase sensitivity only if observers can discriminate the differences in performance that are associated with each scale point.

The number of points on each rating scale was increased from 7 to 10 to improve the sensitivity of the present rating form. However, preliminary work indicated that observers could not discriminate the subtle differences in controller actions using a 10-point rating scale, so an 8-point format was adopted in the final version. The rating form was constructed with labels and statements describing the necessary controller actions for each scale point. This change resulted in a format that is quite different from the SACHA rating form. The present rating form has generic scale point descriptions that are used to make ratings in the different performance areas. The SACHA rating form has unique scale point descriptions (i.e., performance examples) for each performance category.

Several controller actions were identified for each of the performance areas as another design feature of the present rating form. These controller actions were observable behaviors that ATCSs should always look for when evaluating controllers. Many of these controller actions were included in the SACHA rating form as performance examples, and other actions were identified by the ATCSs who did the preliminary work with the present form. The accurate categorization of these controller actions is an important part of the rating form.

The present rating form included a comment section for each of the performance areas. The comment sections were used for describing the effective and ineffective controller actions observed during the simulations. The comments served as a justification for the ratings given and helped the researchers understand the observations that led to each rating. Also, ATCSs used the comment sections to identify any observed controller actions not listed in the rating form that were relevant to ATC performance. The decisions of how much to write and whether to include comments for each performance area were at the discretion of the ATCSs, but they were encouraged to write as much as possible.

The rating form was also revised at the recommendation of the six ATCSs who participated in the present study. During the training session, the ATCSs agreed that the rating scale labels were confusing and not necessary, so the researchers removed the scale labels. Also, the ATCSs recognized the importance of the scale point descriptions and worked to improve the terminology. Although the ATCSs thought the final product was valid and very usable, they suggested further work on the scale point descriptions. The performance category, Managing Multiple Tasks, was omitted because the ATCSs thought that the performance areas and controller actions in this category should be moved to the prioritizing category for better organization. The ATCSs added two new performance areas, ensuring positive control and correcting own Errors in a timely manner, to the maintaining attention and situation awareness category.

2.3 Airspace and Traffic Scenarios

The videotapes used in the present study were recorded during a simulation conducted by Guttman et al. (1995) with Atlantic City International Airport TRACON (ACY) controllers. The purpose of the Guttman et al. study was to develop and validate a generic TRACON (GEN) airspace that would be used as a standard testing environment in future ATC simulations. GEN represented a fictitious airspace designed to provide a realistic environment for controlling traffic and to be relatively easy for controllers to learn. GEN included the elements of a typical terminal sector but had different boundaries, navaids, traffic routes, and operating procedures from the ACY model. The ACY model was originally developed in a simulation study conducted by Sollenberger and Stein (1995) and accurately represented the airspace, traffic, and operations of the controllers' facility.

In the generic sector study, 10 ACY controllers worked 2 days of traffic scenarios using both airspaces. On the first day, controllers worked four training scenarios using GEN that were designed to familiarize the participants with the new airspace. On the second day, controllers worked two scenarios in ACY and two scenarios in GEN. One of the scenarios from each airspace was designed to represent a low volume of traffic and the other scenario represented a high volume of traffic. Low traffic scenarios consisted of 33 or 35 aircraft appearing within the one-hour duration of each scenario. High traffic scenarios consisted of 49 or 50 aircraft appearing within the same one-hour period. Only scenarios presented on the second day were audio and video recorded.

3. Method

3.1 Observers

Six TRACON supervisors and training staff specialists from ATC facilities nationwide participated as observers in this study. All observers were FPL controllers and five had actively controlled traffic in the past year. Four of the observers were from Level 5 facilities, one was from a Level 4 facility, and one was from a Level 3 facility. The controllers had a mean age of 43.5 years.. The observers had from 13 to 28 years of experience (Mean = 18.83, SD = 6.31) as active controllers and from 4 to 19 years of experience (Mean = 10.17, SD = 4.92) in training and evaluating controllers.

3.2 Simulation Facility

The study was conducted in the Research Development and Human Factors Laboratory at the FAA William J. Hughes Technical Center in New Jersey. The laboratory briefing room and video projection system were used to present the videotapes from the generic sector study. Two different views of the simulation were presented on large projection screens. The first view was recorded by a camera located in the corner of the simulation room and showed an over-the-shoulder view of the controller's upper body, workstation equipment, and radar display. In this view, it was not possible to read the writing on flight progress strips or the data on the radar display. However, the controller's head and arm movements and interactions with the workstation equipment were clearly visible. The second view was a graphical playback of the

traffic scenario using the simulation software, ATCoach (UFA Inc., 1992). The playback view showed all the information on the controller's radar display in a large and easily readable format. Both views were simultaneously presented on different screens and synchronized with an audio recording of communications between the controllers and simulation pilots.

3.3 Training

The accuracy of any measurement system depends not only on the measuring device but also on the users of the measuring device. Therefore, a good training program is a necessary component of measurement systems and is essential to the reliability of any observer rating form. The six observers in the present study participated in a training program before using the rating form to make formal evaluations of the videotapes. The training program was conducted in two separate sessions by a research team of psychologists and SMEs. The first training session lasted 1 day and was designed to help the observers learn the airspaces in the simulation. The second training session lasted 3 days and was designed to help the observers become proficient with the rating form.

In the first training session, the observers were informed about the goals of the study, how the study was going to be conducted, and what was expected from them as participants. All aspects of the simulation setup, equipment, software, and data collection capabilities were explained. A modified version of the training manual originally developed in the generic sector study was used to assist observers in learning the airspaces. The training manual described the letters of agreement (LOAs) for both airspaces and included attachments illustrating the sector layouts, arrival and departure routes, transfer-of-control points, and approach plates for all airport runways. The manual was made available for observers to read after the first day and was reviewed during the training session. The first session was concluded with several hands-on training scenarios where observers controlled some light air traffic.

In the second training session, the design process and development work that had been completed on the rating form was explained. Several steps encouraged the observers to adopt mutual evaluation criteria for their ratings. First, the research team discussed common rater biases and how to avoid them. Then, the observers reviewed the rating form and discussed their interpretations of the terminology. Next, the observers used the rating form while viewing six practice tapes. After each tape was viewed, the observers' ratings were displayed on the projection screen for everyone to see, and they began a discussion of what they saw and why they selected their ratings. Each discussion lasted approximately one hour and helped to clarify some of the ambiguities in the rating form and identify the observers whose rating style differed a great deal from the others. Several modifications were made to the rating form by the conclusion of the training program.

3.4 Procedure

The present study was scheduled to be completed within 2 weeks (i.e., 10 work days). The Monday of the first week was reserved for the participants' travel to Atlantic City. The remaining 4 days consisted of training the observers. The first 4 days of the second week were

scheduled for the actual videotape evaluations. The final Friday was reserved for debriefing and the participants' return trip.

On the first day of the study, the controllers completed a Background Questionnaire to obtain information about the group of participants. On the last day, a Final Questionnaire was completed. In the Final Questionnaire, observers provided weighting values that indicated the relative importance of the six performance categories. The weights were used to calculate an overall performance score for the controller in each of the videotapes. Specifically, the weight for each category was multiplied by the mean of the ratings within the category, and the results were added to produce a weighted overall performance score ranging from 1.0 to 8.0. Also, observers ranked the controllers who participated in the simulations. Finally, observers responded to a few questions about the training program and methods used in the present study. Both questionnaires are presented in Appendix C.

Because of time limitations, the entire set of 40 videotapes from the generic sector study could not be viewed. The researchers randomly selected 4 of the 10 controllers who participated in the generic sector study and used all 4 videotapes from each controller. Additionally, the observers viewed one tape from each of the controllers a second time to obtain a measure of reliability on repeated scenarios. In total, the observers viewed 20 one-hour videotapes; 5 tapes were shown on each day.

The presentation order of the videotapes (see Table 1) was selected so that similar tapes were not viewed consecutively, which may have led observers to evaluate each tape comparatively instead of independently. The videotapes were arranged so that only one of the controllers and only one

Table 1. Presentation Order of Videotapes

Evaluation Session	Presentation Order				
Day 1	S2-AH	S3-GL	S1-AL	S4-GH	S2-AL
Day 2	S3-GH	S1-AH	S4-AL	S2-GL	S3-AH
Day 3	S4-AH	S1-GL	S3-AL	S2-GH	S4-GL
Day 4	S1-GH	S2-AH'	S3-GL'	S4-GH'	S1-AL'

Note: S# indicates controller identification code
A and G indicate ACY and GEN, respectively
L and H indicate low and high traffic scenarios, respectively
The apostrophe indicates a videotape that was repeated from the first evaluation day

of the scenarios were shown twice on the same day. The controller who was viewed first on each day was shown working a different scenario on the last tape of the day. Also, the scenario that was shown twice during the day was worked by different controllers on the two occasions and separated in the presentation order. Videotapes from ACY and GEN were alternated in presentation as much as possible. Low and high traffic scenarios were alternated also. The

videotapes that were repeated to assess test-retest reliability were viewed on the first and last evaluation days.

Besides analyzing the ratings obtained from the observers, the present study examined the relationship between the ratings and a subset of SEMs routinely collected in ATC simulation research (Buckley et al., 1983). A list of the SEMs is presented in Table 2.

Table 2. System Effectiveness Measures Recorded During the Simulation

Abbreviation	Description
NCNF	Number of Conflicts (less than 3 nmi and 1,000 ft separation)
NALT	Number of Altitude Assignments
NHDG	Number of Heading Assignments
NSPD	Number of Speed Assignments
NPTT	Number of Push-to-Talk Transmissions
CMAV	Cumulative Average of System Activity/Aircraft Density (number of aircraft within 8 nmi of another aircraft)
DIST	Total Distance Flown by Aircraft
ATWIT	Air Traffic Workload Input Technique Rating

4. Results and Discussion

The primary purpose of this research was to develop and evaluate a rating form intended to assess air traffic controller performance. One of the most important criteria for the successful evaluation of a new rating form is reliability. Two types of reliability were assessed: inter-rater and intra-rater reliability. Inter-rater reliability refers to the consistency of the ratings between the observers. Intra-rater reliability is often called test-retest reliability and refers to the consistency of observer ratings on repeated occasions. Analyses were conducted to examine both inter- and intra-rater reliability.

There were several other issues addressed in the present study. First, the study investigated which SEMs collected during the simulations were good indicators of controller performance. The SEMs are ATC performance measures that have the desirable qualities of an objective, reliable, and automated data collection system. However, many of the performance areas listed in the present rating form could not be easily measured by automated methods. A correlation analysis of these different measurement systems was conducted to identify any SEMs that were related to the observer ratings.

The study also examined the relationship between observer ratings in different performance areas. Each performance category on the rating form consisted of several related performance areas. Therefore, observer ratings from the same performance category should be related. Correlation analyses were conducted to determine if ratings from different performance categories were related and if any ratings were related to overall controller performance. The analyses also determined if the different rating areas were truly measuring independent aspects of controller performance.

Finally, the observer ratings in the present study were used to support the generic sector development research of Guttman et al. (1995). The purpose of the generic sector study was to develop and validate a fictitious airspace that would be used as a standard testing environment in future ATC simulations. To validate the generic sector, controller performance using ACY was related to performance using GEN. Although many ATC measures were collected in the generic sector study, only one observer made over-the-shoulder ratings. In the present study, six observers provided ratings of controllers using the different airspaces. Correlation analyses were conducted to determine the relationship between the ACY and GEN ratings to evaluate the validity of the generic sector.

4.1 Reliability of Observer Ratings

The analysis that was used to calculate inter-rater reliability is based upon analysis of variance (ANOVA) and is more fully discussed in Winer, Brown, and Michels (1991). The analysis provides an estimate of the reliability of a single measurement (i.e., one observer's rating) and represents what is often called the intraclass correlation. The results of the analysis produce a reliability coefficient (or r value) that ranges from 0 to 1.0 and indicates the consistency of the obtained measurements. A coefficient of 1.0 means the measurements are perfectly consistent, and the closer the coefficient is to 0, the more inconsistent the measurements.

A correlation analysis was used to calculate intra-rater reliability. The results of a correlation analysis produce a correlation coefficient (also denoted by r) that ranges from -1.0 to +1.0 and indicates the strength of the relationship between two variables. A coefficient of 0 means that no relationship exists, while -1.0 and +1.0 indicate perfect relationships. A positive coefficient (or direct relationship) means that as the value of one variable increases, the other variable increases. A negative coefficient (or inverse relationship) means that as the value of one variable increases, the other variable decreases. In the context of reliability, the relationship between two measurements indicates the consistency of the measures, and negative coefficients are usually not obtained.

A correlation coefficient is considered to be statistically significant if its absolute magnitude exceeds a given critical value, which depends upon the number of degrees of freedom in the experimental design. Usually, a p value (or significance level) is reported, which represents the probability that the calculated coefficient could exceed the critical value by chance alone.

The inter-rater reliability analysis was based upon a maximum of 120 observations (6 observers times 20 total scenarios). The intra-rater reliability analysis was based upon a maximum of 24 observations (6 observers times 4 repeated scenarios). However, many of the coefficients were

based upon fewer observations because of intentional "not applicable" responses by observers. The critical values associated with 120 and 24 observations are .23 and .52, respectively, at a significance level of $p < .01$.

The results of the inter-rater reliability analysis are reported in Figure 1. The coefficients range from .01 to .90. However, 72% of the ratings exceed .60 and 56% exceed .80. The overall ratings for each performance category are generally more reliable than the individual ratings within the category. The weighted overall performance score is $r = .90$. There are no generally accepted guidelines on the minimum level of acceptable reliability, and a great deal depends on the purpose of the measurement being evaluated (Guilford, 1954, p. 388). Most researchers and practitioners would find a reliability coefficient of $r = .90$ as quite acceptable for research purposes.

The results of the intra-rater reliability analysis are shown in Figure 2. The coefficients range from .43 to .91; 72% of the ratings exceed .60 and 28% exceed .80. The overall ratings for each performance category are generally more reliable than the individual ratings within the category. The weighted overall performance scores have a reliability of $r = .86$, which was close to that achieved for the inter-rater reliability.

One result from the inter-rater reliability analysis was that most of the ratings scales showed moderate to high coefficients. There were some scales that had low reliability. There are at least two factors that may have affected the consistency of observer ratings. The first is establishing a mutual rating standard for each of the performance areas, which refers to specifying when controller performance warrants a 1, 2, or 3 rating, etc. Each of the observers had their own personal standards before participating in the study. These standards have diverged more in some rating areas than others. Some may have begun as more strict evaluators, while others may have been more lenient.

The purpose of the training program was to help establish a set of mutual standards that everyone understood and was comfortable in using. The scales within the instrument demonstrated a range of reliability coefficients. This variability had to be the result of either the training and or the rating scales themselves. Observers may have been reluctant to completely abandon their own personal criteria in favor of the standards identified by the entire group. The low reliability of a few of the rating scales may have been due in part to a lack of understanding or compliance with the agreed-upon standards for those scales.

The second factor that may have affected the consistency of observer ratings was defining observable controller actions that were unambiguous and easily detected by experienced ATCSs. The rating form was designed to provide a list of observable controller actions for each performance area. Some of the ratings may have required observers to make inferences about the controller's thinking or plans when the actions were completed. Other ratings may have required observers to detect controller actions that were easily overlooked during busy conditions.

The overall category ratings were more reliable than the individual ratings within each category, and the weighted overall performance scores were usually the most reliable single measure of controller performance. Essentially, this means that observers found it easier to agree on the

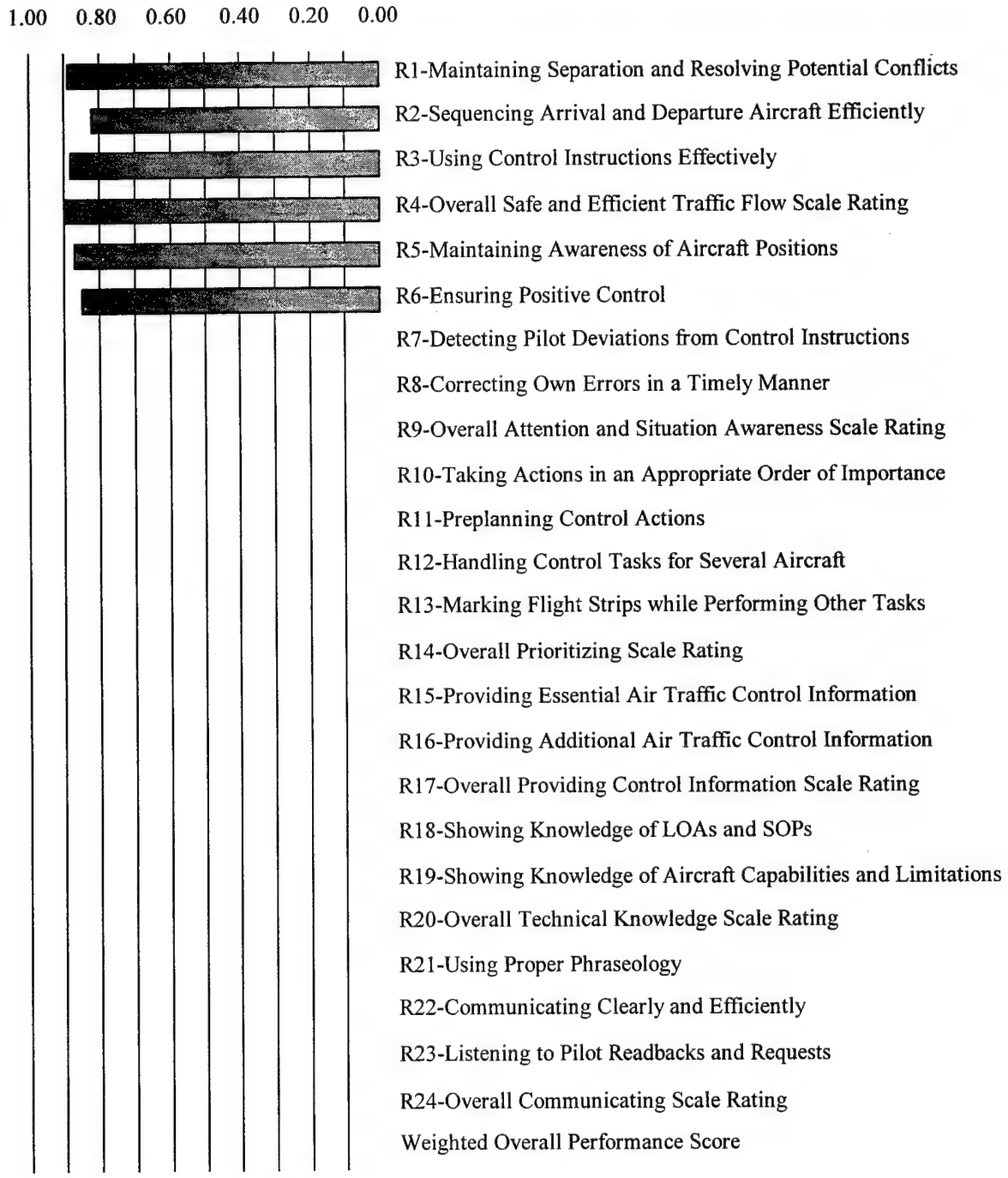

Figure 1. Inter-rater reliability for each of the performance areas using intraclass correlations.

Note: LOAs are letters of agreement between facilities and SOPs are standard operating procedures.

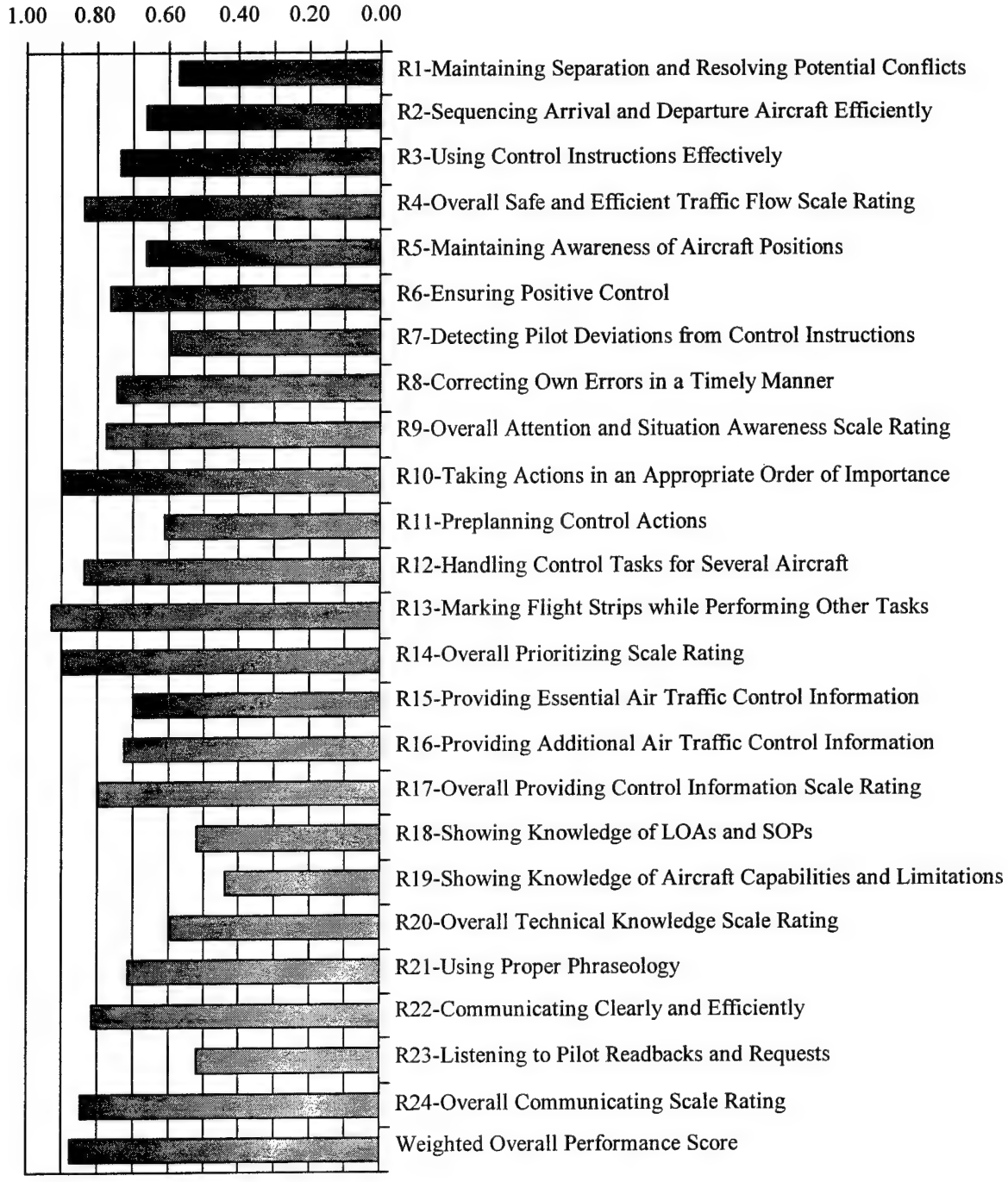

Figure 2. Intra-rater reliability for each of the performance areas using Pearson correlations between repeated scenarios.

general aspects of controller performance than on the details of performance. The reasons for the differential consistency of observer ratings may be related to the two factors discussed previously. Also, because the weighted overall performance scores were based upon all six category ratings, it is reasonable that this measure would be the most reliable.

The results of the intra-rater reliability analysis were similar to the inter-rater reliability analysis. There are many reasons why observers might change their ratings when viewing the same videotape on two occasions. After viewing several videotapes, observers probably gained a better understanding of the skill level of the controllers who participated in the simulations. This experience could have induced some shifting in rating standards and made the observers more strict or lenient on the second viewing. Also, observers had seen each controller in four videotapes before rating the repeated scenarios. If their impression of the controller's skill level changed, this might influence their final ratings. Finally, as observers gained more experience with the new rating form, they may have improved their observation skills and were able to detect controller actions more easily.

4.2 Relationship Between Observer Ratings and System Effectiveness Measures

In testing and evaluation research at the FAA William J. Hughes Technical Center, ATCSs routinely serve as over-the-shoulder observers. Researchers depend on the judgments of these experts to determine if a proposed change to the system has any negative consequences for controllers in the performance of their jobs. Expert opinions have validity as a measurement system, because researchers know that experts have the training, experience, and qualifications to evaluate controller performance. However, experts often disagree, make errors, and are influenced by their own subjective biases. Therefore, reliability is always a concern for any measurement system relying on expert opinions. For this reason, computers are used whenever possible to collect objective performance measures in simulations. However, the main concern with automated measurement systems is that they are unable to capture the subtle aspects of controller style that are the essence of controller performance. In practice, using both experts and computers to evaluate controller performance is the best method to ensure quality in testing and evaluation research.

One method for determining which SEMs were good indicators of controller performance is to determine the relationship between the SEMs and the observer ratings. The correlation analysis examining the relationship between the SEMs and the weighted overall performance scores was conducted for this purpose. The correlations in Figure 3 show the relationship between the SEMs and the weighted overall performance scores that were computed from the observer ratings. The figure indicates a negative (or inverse) relationship between all eight SEMs and the performance scores. In general, the correlations ranged from -.03 to -.63.

The SEMs indicated the frequency of controller actions that were necessary to control the traffic. The negative correlations were in the expected direction for this set of measurements and indicated that fewer controller actions were rated more favorably by the observers. NPTT (see Figure 3) had the strongest correlation with the observer ratings and suggests that controller transmissions were the best indicator of overall performance. NSPD had the weakest correlation and suggests that speed modifications may have been seen by the observers as less important indicators of controller performance. However, the controllers in the present study rarely attempted speed control, which is more commonly used at major terminal facilities. Given that there was little variance in the speed control SEM, a low correlation may have, in fact, simply been the result of the lack of variance in one variable.

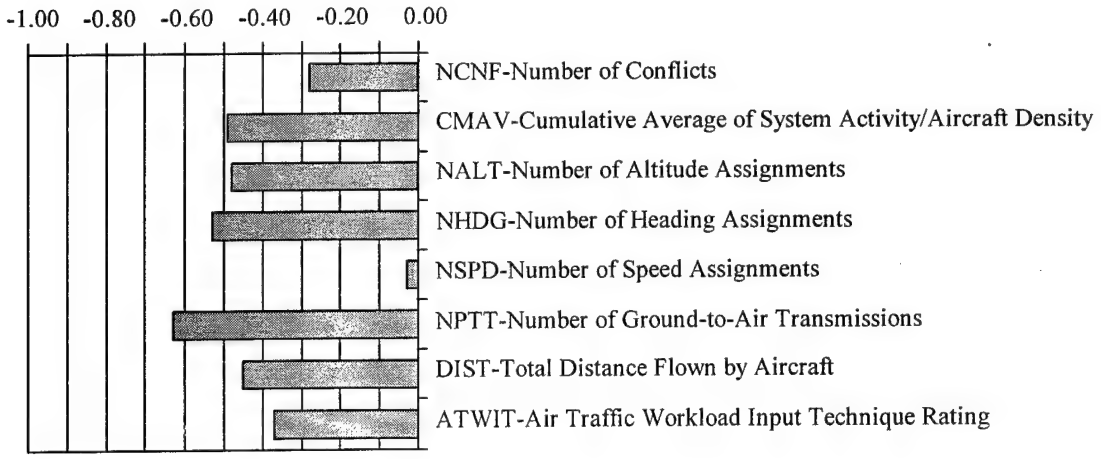

Figure 3. Correlations between the system effectiveness measures and the weighted overall performance scores.

4.3 Relationship Between Observer Ratings in Different Performance Areas

The results of the correlation analysis examining the relationships between the observer ratings in different performance areas are summarized in Table 3 and Figure 4 (see Appendix D for the complete correlation matrix). Table 3 shows the correlations between the overall ratings of the six performance categories. The table indicates correlation among the rating scales ranging from .55 to .89. Figure 4 shows the correlations between the individual performance areas and the overall weighted performance scores. The figure indicates correlations for many of the rating scales, especially the overall category ratings. The coefficients range from .47 to .94; 88% of the ratings exceed .60, and 50% of the ratings exceed .80.

Table 3. Correlations Between the Observer Ratings of the Major Performance Categories

	R4	R9	R14	R17	R20	R24
R4-Overall Safe and Efficient Traffic Flow Scale						
R9-Overall Attention and Situation Awareness Scale Rating	0.89					
R14-Overall Prioritizing Scale Rating	0.84	0.83				
R17-Overall Providing Control Information Scale Rating	0.63	0.55	0.57			
R20-Overall Technical Knowledge	0.69	0.62	0.57	0.56		
R24-Overall Communicating Scale Rating	0.71	0.67	0.75	0.61	0.56	

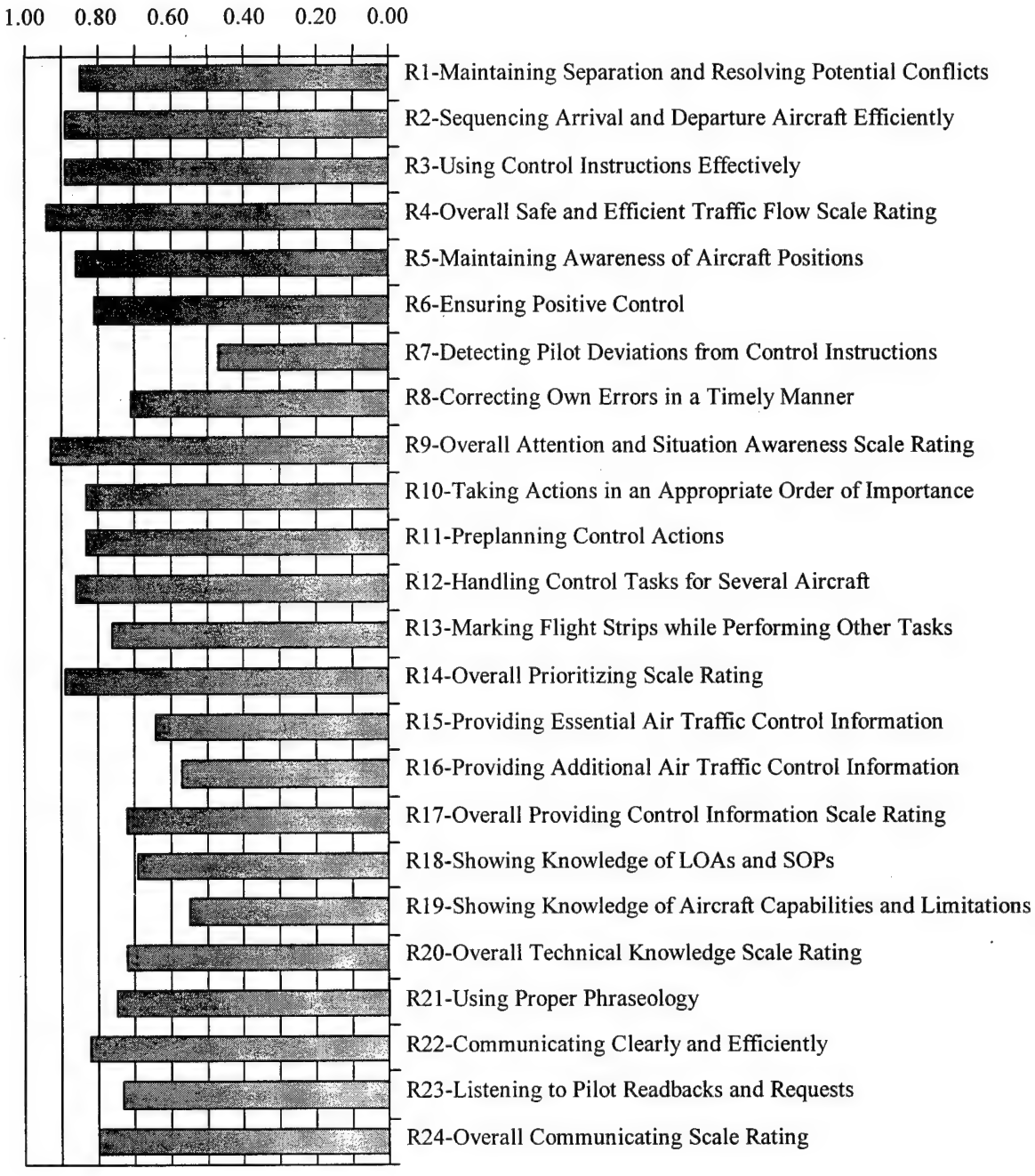

Figure 4. Correlations between the observer ratings and the weighted overall performance scores.

As expected, all six overall performance categories were related to some degree. There are several possible explanations for the relationship among these first three performance categories. Perhaps, good situation awareness led to good prioritizing and safe and efficient traffic flow. Perhaps, good prioritizing led to good situation awareness and safe and efficient traffic flow. Alternatively, there may not be a causal relation at all, and another factor may have determined performance such as the raters' inability to separate the various dimensions. The last three dimensions were not as highly correlated with each other as were the first three. It is possible that some controllers may forget to provide control information but have good technical knowledge and communications skills and demonstrate good skills in all other performance categories. This is admittedly a speculation on the thought processes of the observer raters. Appendix D provides a summary of the correlations that evaluated the inter-relationships between all the variables in the rating form. As expected, most of the performance areas were at least moderately related to the weighted overall performance score. However, the overall ratings in maintaining safe and efficient traffic flow, maintaining attention and situation awareness, and prioritizing showed the strongest relationships. Also, the individual ratings within these performance categories were very strongly related to overall performance, except for detecting pilot deviations from control instructions. It stands to reason that the individual ratings within these first three performance categories had larger correlations because the categories were assigned the largest weights by the observers. On the other hand, each rating has such a small mathematical contribution to the overall weighted performance score that this cannot be the only reason for the correlations. Additionally, the overall category ratings were not included in the calculation of the weighted overall performance scores, and these correlations were high also. Ideally, in a performance measurement environment, researchers strive for a list of variables that are reasonably independent. However, in practice, this is rarely achieved, and the variables that enter into a rating process are often a compromise. This compromise exists between independence and the achievement of some sort of face validity. Users and sponsors need to feel comfortable that the performance dimensions important to them are adequately covered. The table in appendix D shows that there is redundancy in the rating form and process. Ironically one of the least redundant scales which correlates lowest with the other scales was R-7, Detecting Pilot Deviations From Control Instructions. This scale was also the least reliable in the whole process and is being considered for deletion on subsequent versions of the form. Given a choice between reliability of the scale and the redundancy against other scales, reliability will win out in the long run and a certain degree of redundancy will be accepted.

4.4 Relationship Between Observer Ratings in Different Airspaces

The results of the correlation analysis examining the relationship between ACY and GEN are reported in Figure 5. These results represent pooling across scenarios observed. The coefficients range from .14 to .80, and 40% of the ratings exceed .60. The overall ratings for each performance category generally show stronger relationships than the individual ratings within the category with one exception, R-14, which was a little lower than one of its subscales. The reader will recall that the overall ratings were actually made by the observer and did not represent a mathematical composite of the other scales within the category. The weighted overall performance score correlated $r = .77$ between the two airspaces.

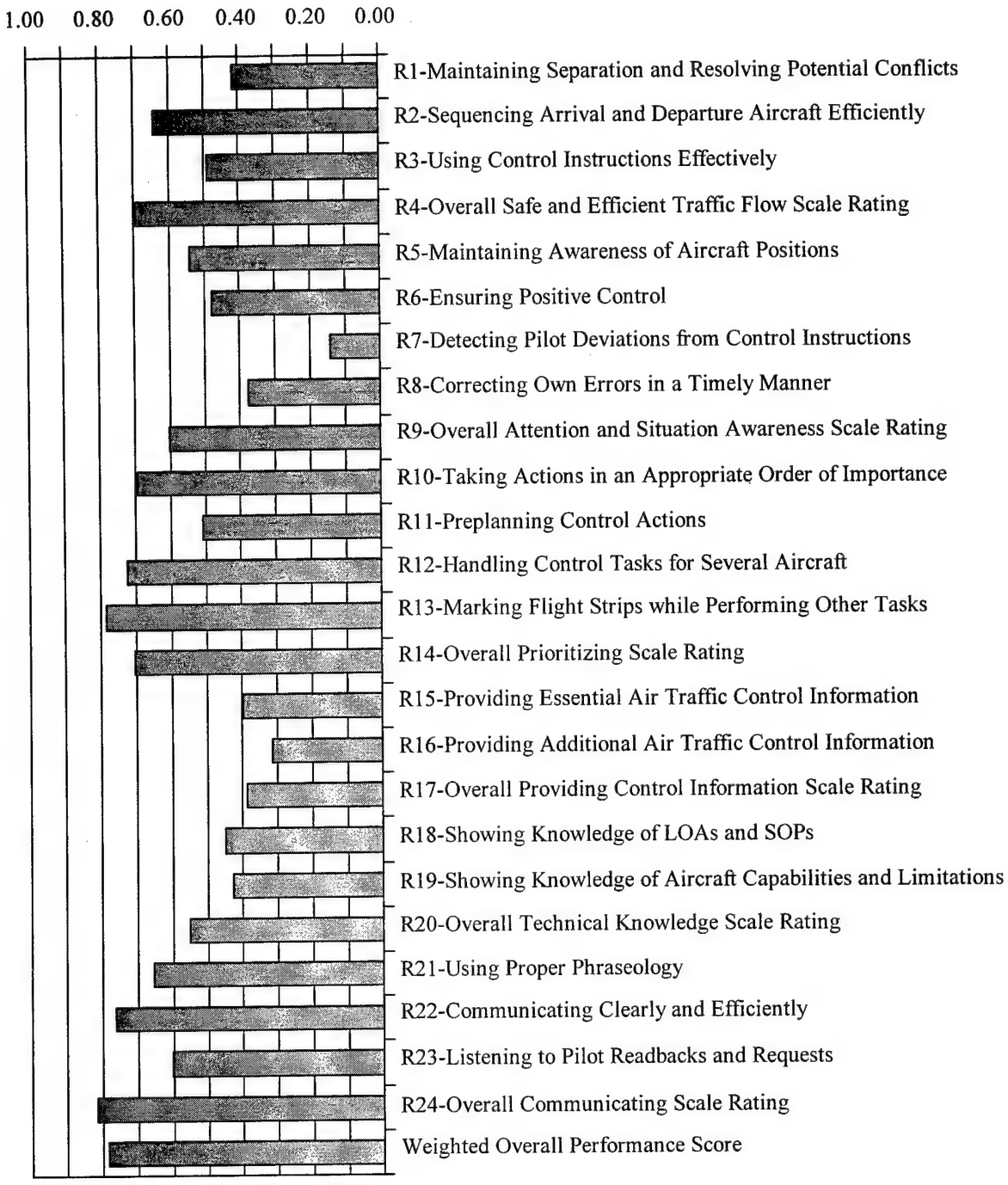

Figure 5. Correlations between the observer ratings of controllers using ACY and GEN.

In simulation experiments, a large sample of controllers from different ATC facilities is desirable to ensure that the results of a study are generalizable to the entire controller population. Also, selecting controllers from different facilities is often necessary because any single facility is unlikely to be able to spare the large number of staff needed for a research assignment.

However, using controllers from different facilities creates the problem of finding a standard testing environment where the airspace, traffic, and operating procedures are equally familiar to all controllers. The solution to this problem is to develop and validate a generic sector. The main requirements for a generic sector are that the airspace be a realistic environment for controlling traffic and be relatively easy for controllers to learn.

In a realistic generic sector, controller performance using their own airspace should be comparable to their performance in the generic airspace. The correlation analysis using the observer ratings from ACY and GEN was conducted to examine this relationship. Although the positive correlations were in the appropriate direction, the relationships were not as strong as expected. A few of the observer ratings, especially in the prioritizing and communicating areas, showed good relationships. Essentially, this means that GEN involved prioritizing and communicating controller actions similar to ACY. On the other hand, the results suggest that the requirements in the providing control information and technical knowledge areas were different when using ACY. These differences could have occurred because of different LOAs, standard operating procedures (SOPs), aircraft types, or traffic situations in the two airspaces.

4.5 Summary of Final Questionnaire

Table 4 displays the mean weights observers assigned to indicate the relative importance of the six performance categories. As shown, the observers generally assigned the highest weights to maintaining safe and efficient traffic flow and maintaining attention and situation awareness. They assigned the lowest weights to providing control information, technical knowledge, and prioritizing. Communicating received the third highest mean priority weighting, although it had the highest variability. The fact that there was complete agreement concerning the relative priority of providing control information is notable. These weights were used to compute the final weighted performance evaluation scores from each observer for each run.

Table 4. Mean Observer Weights Assigned to Each Performance Category

	Maintaining Safe and Efficient Traffic Flow	Maintaining Attention and Situation Awareness	Prioritizing	Providing Control Information	Technical Knowledge	Communicating
Means	27.50	23.33	11.67	10.00	11.67	15.83
Std	2.5	3.7	2.3	0	3.7	4.5

The weighted overall performance scores provided one composite data point for each of the four observed controllers that appeared on the videotape. Researchers used these data as basis for ranking the four controllers based on the input of each observer for Atlantic City, Generic, and Combined airspace under which the observed controllers performed. Table 5 provides the Spearman rank order correlations under the three airspace conditions.

Table 5. Spearman Inter-Rater Correlations

Atlantic City Airspace

Rater

	1	2	3	4	5	6
1		.8	.8	.8	.8	1.0
2			1.0	1.0	1.0	1.0
3				1.0	1.0	1.0
4					1.0	1.0
5						1.0

Generic Airspace

Rater

	1	2	3	4	5	6
1		.8	1.0	1.0	1.0	.8
2			.8	.8	.8	1.0
3				1.0	1.0	.8
4					1.0	.8
5						.8

All Airspace Combined

Rater

	1	2	3	4	5	6
1		.8	1.0	1.0	1.0	1.0
2			.8	.8	.8	.8
3				1.0	1.0	1.0
4					1.0	1.0
5						1.0

Table 5 summarizes the inter-rater reliability if the overall weighted performance scores are used to rank order the four observed controllers. This was a small sample study and the correlations reported were interesting in that they took only two values $r = .8$ and $r = 1.0$. With only four controllers being observed, this meant entering a correlation significance table with very few degrees of freedom. Given that the .8 correlations were not significant from zero, those that were 1.0 were significant. A cursory examination of Table 5 indicates that for both the Atlantic City airspace and the combined data for all the airspace observed, the majority of the relationships were $r = 1.0$. When rank ordered based on the weighted performance scores, the observed controllers fell into relatively the same order for the 6 raters. The observers were also satisfied with the quality of what they observed and the training they received to use the rating form.

A summary of the observer responses to questions about the training program and videotape methodology used in the study are shown in Table 6. The research team discussed the training and methodology with the observers many times during the study. The observers agreed that the training was very good, and there were no problems with the methods employed. The summary of the questionnaire responses confirmed these comments and provided a method to quantify the opinions expressed by the observers.

Table 6. Observer Responses to Questions on the Final Questionnaire

Question	Mean	SD
1. As compared to viewing controllers "live," the videotapes showed sufficient information for me to make my evaluations.	7.33	2.66
2. The training period was sufficient for me to become familiar with the new rating form.	8.67	1.03

5. Conclusions

This was a small sample study designed to serve several purposes. The goal of evaluating the performance rating form and accompanying training package was first. The second purpose was to determine the feasibility of using videotape and simulation play back capabilities as a source of stimuli for the observer raters. These goals were achieved although not perfectly.

The videotape and playback capabilities functioned and served their purpose during the study. However, it took a considerable amount of technical help to keep the system on track and to maintain the realism of the images and sounds that the observers received. Fortunately, most of this effort was behind the scenes, and the observers saw the system running smoothly and efficiently. They commented on the quality of the playbacks and were willing and able to respond with complete ratings backed up by more notes than they had ever taken before. This was part of the training program.

Reliability of the rating scales varied across a range with some scales, such as R-7, being so low as to be questionable in terms of the quality of measurement. However, most of the reliability on individual scales were in the $r = .7$ to $r = .9$ range with the summary scales for each performance

area generally running higher than the individual scales. When observers priorities were taken into consideration and the overall weighted composite scores were computed, reliability using intraclass correlations and rank orders of the observed controllers was, for the most part, acceptable. This is an admittedly arbitrary conclusion in that there is no finite standard for acceptable reliability, and it depends on the conditions under which the reliability is computed and for what purpose.

There is currently no reliability data available on any controller performance rating form in use today. The data for this new research-oriented form is all there is. This study, which was based on the observation of controllers performing in a TRACON environment, will be followed by another research effort using supervisory controllers from en route centers observing controllers who have worked simulated center airspace. Then, we will have even more reliability data and will carry out the next study having learned from the work reported here.

The study successfully demonstrated the feasibility of using videotape presentation in testing and evaluation research. This technique represents a cost-effective method for obtaining evaluations from a large number of observers. Expenses can be greatly reduced by having the research team travel with the equipment and tapes to ATC facilities nationwide instead of paying for the travel and per diem costs of the observers. Also, there is always some concern that the presence of an observer (or several observers) making over-the-shoulder evaluations affects controller performance. The videotape method avoided these potential problems because the small cameras were much more unobtrusive than observers standing behind the controllers and writing notes on clipboards.

References

Anastasi, A. (1988). *Psychological testing*. New York: Macmillan.

Bailey, R. W. (1982). *Human performance engineering: A guide for system designers*. Englewood Cliffs, NJ: Prentice Hall.

Bass, B. M., & Barrett, G. V. (1981). *People, work and organizations*. Boston: Allyn and Bacon.

Bender, H. E., Eichel, E., & Bender, J. (1985). Performance appraisal: An area for HF interventions. In H. W. Hendrick and O. Brown (Eds.), *Human factors in organizational design and management*. Amsterdam: Elsevier.

Boone, J. O., & Steen, J. A. (1981). A comparison between over the shoulder and computer derived measurement procedures in assessing student performance in radar air traffic control. *Aviation, Space and Environmental Medicine*, *52*(10), 589–593.

Buckley, E. P., DeBaryshe, B. D., Hitchner, N., & Kohn, P. (1983). *Methods and measurements in real-time air traffic control system simulation* (DOT/FAA/CT-TN83/26). Atlantic City, NJ: DOT/FAA Technical Center.

Buckley, E. P., O'Connor, W. F., Beebe, T., Adams, W., & MacDonald G. (1969). *A comparative analysis of individual and system performance indices for the air traffic control system* (FAA-NA-69-40). Atlantic City, NJ: National Aviation Facilities Experimental Center. (NTIS No. AD-710-795).

Endsley, M. R., & Rodgers, M. D. (1994). Situation awareness information requirements analysis for en route air traffic control. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting* (pp. 73–75). Santa Monica, CA: Human Factors and Ergonomics Society.

Federal Aviation Administration (1988). *Profile of operational errors in the national airspace system calendar year 1987*. Washington, DC: Department of Transportation, Federal Aviation Administration.

Federal Aviation Administration (1995). *Administrator's fact book* (AMS-400). Washington, DC: Department of Transportation, Federal Aviation Administration.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw Hill.

Guttman, J., Stein, E. S., & Gromelski, S. (1995). *The influence of generic airspace on air traffic controller performance*. (DOT/FAA/CT-TN95/38). Atlantic City, NJ: DOT/FAA Technical Center.

Hedge, J. W., Borman, W. C., Hanson, M. A., Carter, G. W., & Nelson, L. C. (1993). *Progress toward development of ATCS performance criterion measure* (Institute Report No. 235). Minneapolis: Personnel Decisions Research Institutes, Inc.

Hennessy, R. T. (1990). Practical human performance testing. In H. R. Booher (Ed.), *Manprint* (pp. 433–470). New York: Van Nostrand.

Hopkin, V. D. (1980). The measurement of the air traffic controller. *Human Factors*, *22*(5), 547–560.

Hopkin, V. D. (1991). The impact of automation on air traffic control systems. In J. A. Wise, V. D. Hopkin, and M. L. Smith (Eds.), *Automation and systems issues in air traffic control* (pp. 3–19). Berlin: Springer-Verlag.

Kinney, G. C., Spahn, M. J., & Amato, R. A. (1977). *The human element in air traffic control: Observations and analysis of the performance of controllers and supervisors in providing ATC separation services* (MTR-7655). McLean, VA: The MITRE Corporation.

Robertson, A., Grossberg, M., & Richards, J. (1979). *Validation of air traffic controller workload models* (FAA-RD-79-83). Cambridge, MA: US Department of Transportation.

Rodgers, M. D. (1993). *An examination of the operational error data base for air traffic control centers*. (DOT/FAA/AM-TN93/22). Oklahoma City, OK: FAA Civil Aeromedical Institute.

Seamster, T. L., Redding, R. E., Cannon, J. R., Ryder, J. M., & Purcell, J. A. (1993). Cognitive task analysis of expertise in air traffic control. *The "international Journal of Aviation psychology*, *3*(4), b257–283.

Sollenberger, R. L., & Stein, E. S. (1995). *The effects of structured arrival and departure procedures on TRACON air traffic controller memory and situational awareness* (DOT/FAA/CT-TN95/27). Atlantic City, NJ: DOT/FAA Technical Center.

Stein, E. S. (1984a). *The measurement of pilot performance—a master-journeyman approach* (DOT/FAA/CT-TN83/15). Atlantic City, NJ: DOT/FAA Technical Center.

Stein, E. S. (1984b). Observer rating of air traffic controller workload during simulation. In V. Amico and A. B. Clymer (Eds.), *Proceedings of the SCS Simulators Conference*, *14*(1), 288–290.

Stein, E. S. (1984c). The advantages of simulation for the study of air traffic controller workload—Automated measurement techniques. In W. Wade (Ed.), *Proceedings of the 1984 Summer Computer Simulation Conference* (pp. 1174–1178). La Jolla, CA: Society for Computer Simulation.

Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe* (DOT/FAA/CT-TN84/24). Atlantic City, NJ: DOT/FAA Technical Center.

Stein, E. S. (1989). *Parallel approach separation and controller performance* (DOT/FAA/CT-TN89/50). Atlantic City, NJ: DOT/FAA Technical Center.

Stein, E. S., & Buckley, E. P. (1992). *Simulation variables*. Unpublished manuscript.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

UFA, Inc. (1992). *ATCoach* [Computer software]. Lexington, MA: UFA, Inc.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design, 3rd edition*. New York: McGraw-Hill, Inc.

Appendix A
Observer Rating Form

Observer Code _____ Date _____

Controller 1 2 3 4 Sector ACY GEN Traffic LO HI

INSTRUCTIONS

This form was designed to be used by instructor certified air traffic control specialists to evaluate the effectiveness of controllers working in simulation environments. Observers will rate the effectiveness of controllers in several different performance areas using the scale shown below. When making your ratings, please try to use the entire scale range as much as possible. You are encouraged to write down observations and you may make preliminary ratings during the course of the scenario. However, we recommend that you wait until the scenario is finished before making your final ratings. The observations you make do not need to be restricted to the performance areas covered in this form and may include other areas that you think are important. Also, please write down any comments that may improve this evaluation form. Your identity will remain anonymous, so do not write your name on the form. Instead, your data will be identified by an observer code known only to yourself and the researchers conducting this study.

Rating	Scale Point Description
1	Controller demonstrated *extremely* poor judgment in making control decisions and *very* frequently made errors
2	Controller demonstrated poor judgment in making some control decisions and occasionally made errors
3	Controller made questionable control decisions using poor control techniques which led to restricting the normal traffic flow
4	Controller demonstrated the ability to keep aircraft separated but used spacing and separation criteria which was excessive
5	Controller demonstrated *adequate* judgment in making control decisions
6	Controller demonstrated *good* judgment in making control decisions using efficient control techniques
7	Controller *frequently* demonstrated *excellent* judgment in making control decisions using extremely good control techniques
8	Controller *always* demonstrated excellent judgment in making even the most difficult control decisions while using outstanding control techniques
NA	Not Applicable - There was not an opportunity to observe performance in this particular area during the simulation

MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW

1. Maintaining Separation and Resolving Potential Conflicts ..... 1 2 3 4 5 6 7 8 NA

   - using control instructions that maintain safe aircraft separation
   - detecting and resolving impending conflicts early

2. Sequencing Arrival and Departure Aircraft Efficiently ..... 1 2 3 4 5 6 7 8 NA

   - using efficient and orderly spacing techniques for arrival and departure aircraft
   - maintaining safe arrival and departure intervals that minimize delays

3. Using Control Instructions Effectively ..... 1 2 3 4 5 6 7 8 NA

   - providing accurate navigational assistance to pilots
   - avoiding clearances that result in the need for additional instructions to handle aircraft completely
   - avoiding excessive vectoring or over-controlling

4. Overall Safe and Efficient Traffic Flow Scale Rating ..... 1 2 3 4 5 6 7 8 NA

MAINTAINING ATTENTION AND SITUATION AWARENESS

5. Maintaining Awareness of Aircraft Positions ..... 1 2 3 4 5 6 7 8 NA

   - avoiding fixation on one area of the radar scope when other areas need attention
   - using scanning patterns that monitor all aircraft on the radar scope

6. Ensuring Positive Control ..... 1 2 3 4 5 6 7 8 NA

7. Detecting Pilot Deviations from Control Instructions ..... 1 2 3 4 5 6 7 8 NA

   - ensuring that pilots follow assigned clearances correctly
   - correcting pilot deviations in a timely manner

8. Correcting Own Errors in a Timely Manner ..... 1 2 3 4 5 6 7 8 NA

9. Overall Attention and Situation Awareness Scale Rating ..... 1 2 3 4 5 6 7 8 NA

PRIORITIZING

10. Taking Actions in an Appropriate Order of Importance ..... 1 2 3 4 5 6 7 8 NA

- resolving situations that need immediate attention before handling low priority tasks
- issuing control instructions in a prioritized, structured, and timely manner

11. Preplanning Control Actions ..... 1 2 3 4 5 6 7 8 NA

- scanning adjacent sectors to plan for inbound traffic
- studying pending flight strips in bay

12. Handling Control Tasks for Several Aircraft ..... 1 2 3 4 5 6 7 8 NA

- shifting control tasks between several aircraft when necessary
- avoiding delays in communications while thinking or planning control actions

13. Marking Flight Strips while Performing Other Tasks ..... 1 2 3 4 5 6 7 8 NA

- marking flight strips accurately while talking or performing other tasks
- keeping flight strips current

14. Overall Prioritizing Scale Rating ..... 1 2 3 4 5 6 7 8 NA

PROVIDING CONTROL INFORMATION

15. Providing Essential Air Traffic Control Information ..... 1 2 3 4 5 6 7 8 NA

- providing mandatory services and advisories to pilots in a timely manner
- exchanging essential information

16. Providing Additional Air Traffic Control Information ..... 1 2 3 4 5 6 7 8 NA

- providing additional services when workload is not a factor
- exchanging additional information

17. Overall Providing Control Information Scale Rating ..... 1 2 3 4 5 6 7 8 NA

TECHNICAL KNOWLEDGE

18. Showing Knowledge of LOAs and SOPs ..... 1 2 3 4 5 6 7 8 NA

- controlling traffic as depicted in current LOAs and SOPs
- performing handoff procedures correctly

19. Showing Knowledge of Aircraft Capabilities and Limitations ..... 1 2 3 4 5 6 7 8 NA

- avoiding clearances that are beyond aircraft performance parameters
- recognizing the need for speed restrictions and wake turbulence separation

20. Overall Technical Knowledge Scale Rating ..... 1 2 3 4 5 6 7 8 NA

COMMUNICATING

21. Using Proper Phraseology ..... 1 2 3 4 5 6 7 8 NA

- using words and phrases specified in ATP 7110.65
- using ATP phraseology that is appropriate for the situation
- avoiding the use of excessive verbiage

22. Communicating Clearly and Efficiently ..... 1 2 3 4 5 6 7 8 NA

- speaking at the proper volume and rate for pilots to understand
- speaking fluently while scanning or performing other tasks
- clearance delivery is complete, correct and timely
- providing complete information in each clearance

23. Listening to Pilot Readbacks and Requests ..... 1 2 3 4 5 6 7 8 NA

- correcting pilot readback errors
- acknowledging pilot or other controller requests promptly
- processing requests correctly in a timely manner

24. Overall Communicating Scale Rating ..... 1 2 3 4 5 6 7 8 NA

MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW

1. Maintaining Separation and Resolving Potential Conflicts

2. Sequencing Arrival and Departure Aircraft Efficiently

3. Using Control Instructions Effectively

4. Other Actions Observed in Safe and Efficient Traffic Flow

MAINTAINING ATTENTION AND SITUATION AWARENESS

5. Maintaining Awareness of Aircraft Positions

6. Ensuring Positive Control

7. Detecting Pilot Deviations from Control Instructions

8. Correcting Own Errors in a Timely Manner

9. Other Actions Observed in Attention and Situation Awareness

PRIORITIZING

10. Taking Actions in an Appropriate Order of Importance

11. Preplanning Control Actions

12. Handling Control Tasks for Several Aircraft

13. Marking Flight Strips while Performing Other Tasks

14. Other Actions Observed in Prioritizing

PROVIDING CONTROL INFORMATION

15. Providing Essential Air Traffic Control Information

16. Providing Additional Air Traffic Control Information

17. Other Actions Observed in Providing Control Information

TECHNICAL KNOWLEDGE

18. Showing Knowledge of LOAs and SOPs

19. Showing Knowledge of Aircraft Capabilities and Limitations

20. Other Actions Observed in Technical Knowledge

COMMUNICATING

21. Using Proper Phraseology

22. Communicating Clearly and Efficiently

23. Listening to Pilot Readbacks and Requests

24. Other Actions Observed in Communicating

Appendix B
SACHA Rating Form

COMMUNICATING AND INFORMING

Uses clear concise accurate language to get message across unambiguously, talking only when necessary and appropriate; employing proper phraseology to ensure accurate communication; notifying pilots/controllers/other personnel of information that might affect them as appropriate; issuing advisories and alerts to appropriate parties; listening carefully to requests and instructions and ensuring that they are understood; attending to readbacks and ensuring that they are accurate.

Is consistently too wordy, imprecise in phraseology, or uses slang inappropriately during transmissions to pilots and other controllers	Radio and interphone communications are usually easy to understand; at times, may be somewhat wordy or use ambiguous phraseology on the air	Always uses clear, concise phraseology when talking to pilots or other controllers; is very easy to understand
Is careless about informing pilots concerning circumstances that affect them such as weather, nearby traffic etc.	Is normally good at informing pilots about situations and conditions that affect them (e.g., safety related items)	Consistently provides pilots with the information they need such as timely safety alerts, weather advisories, warnings about unpublished obstructions
Often fails to ensure that own instructions are understood; is not very good at picking up on errors in pilot readbacks of clearances, course changes, etc.	For the most part checks to be certain that own instructions are understood; only occasionally fails to pick up on inaccurate readbacks from pilots	Always ensures that own instructions are clearly understood; pays careful attention to pilot readbacks of clearances

1	2	3	4	5	6	7

MANAGING MULTIPLE TASKS

Keeping track of a large number of aircraft/events at one time; conducting two or more tasks simultaneously; remembering and keeping track of aircraft and their positions; remembering what you were doing after an interruption; returning to what you were doing after an interruption and following through; providing pilots with additional services as time allows.

Has difficulty keeping track of several aircraft at the same time; may focus too narrowly on some aircraft while ignoring others	Keeps on top of movement of several aircraft simultaneously while also dealing with routine communication; when very busy may have to simplify the situation to reduce the number of things attended to	Is extremely adept at keeping track of many aircraft while at the same time handling pilot communications, strip work, etc.
Is ineffective at performing multiple tasks simultaneously; prefers to take one thing at a time	Is good at performing two or sometimes more routine tasks at the same time (e.g., monitoring the screen, talking with pilots and handling strips)	Is fully capable of performing two or more complex tasks simultaneously
Interruptions and distractions often cause him/her to forget about some of the immediate air traffic problems; may be slow in recalling what he/she intended to do before the interruption	After an interruption, can usually handle the air traffic problems remaining from prior to the interruption successfully	After an interruption, always quickly remembers where aircraft are or should be, what he or she was doing with the traffic before the interruption, and the intended control strategy

1	2	3	4	5	6	7

TECHNICAL KNOWLEDGE

Knowing the equipment and its capabilities and using it effectively; knowing aircraft capabilities and limitations (e.g., speed, wake turbulence requirements) and using that knowledge; keeping up-to-date on letters of agreement, changes in procedures, regulations, etc.; keeping up-to-date on seldom used procedures or skills.

At times, may not remain current on new letters of agreement, revised air traffic procedures, etc.	Is usually knowledgeable about and up-to-date on all information relevant to controlling traffic (e.g., letters of agreement, air traffic procedures, etc.)	Always keeps up-to-date on letters of agreement, all pertinent procedures and policies, any sector-specific changes (e.g., revised boundaries)
Has basic knowledge of most aircraft's' capabilities, but may make errors related to not knowing aircraft limitations	Has good knowledge of different aircraft capabilities and applies that knowledge to avoid most errors associated with not knowing aircraft limitations	Has thorough knowledge of different aircraft capabilities and as a result never makes errors such as climbing an aircraft beyond its limits, making an inappropriate speed assignment, or requiring an impossibly tight turn
May be unfamiliar with some of his/her equipment and how it works	Is reasonably familiar with his/her equipment and how it works	Is extremely knowledgeable about and familiar with his/her equipment and how it functions

1	2	3	4	5	6	7

REACTING TO STRESS

Remaining calm and cool under stressful situations; handling stressful air traffic conditions in a professional manner.

Becomes shaken and ineffective in emergency situations	Remains calm and cool to most emergency situations	Remains very calm and cool and reacts effectively even in very serious emergency situations such as aircraft inflight emergencies, lost pilots, etc.
Reacts poorly and performance suffers under stressful air traffic conditions	Stays, calm, focused and functional under busy conditions; may be somewhat less effective in very stressful air traffic situations	Stays calm, focused and very functional in busy and very stressful conditions
Does not function effectively when equipment/system problems arise	Shows professional cool in handling routine equipment/system problems	Handles even serious equipment/system degradation problems with professional cool

1	2	3	4	5	6	7

MAINTAINING ATTENTION AND VIGILANCE

Scanning properly for air traffic events, situations, potential problems, etc.; keeping track of equipment and weather status; identifying unusual events and improper positioning of aircraft; recognizing when aircraft have potential for loss of separation; verifying visually that control instructions are followed; remaining vigilant during slow periods.

Has a tendency to focus too narrowly on one air traffic problem and sometimes fails to recognize other potential problems with conflictions, traffic flow, etc.	For the most part, properly scans the scope and monitors aircraft to maintain awareness of air traffic events, potential problems, etc.	Consistently recognizes potentially dangerous conditions such as errors made by pilots (e.g., wrong turns, descending through assigned altitude)
Often does not recognize that an actions is required; is often lax in watching the radar scope and tends to significantly reduce vigilance during slow periods	Is attentive to the radar scope and maintains vigilance, especially during rush periods; may sometimes be inattentive when traffic is light	Always checks and verifies that clearances and other instructions to pilots are followed; remains highly vigilant even during slow periods
Has problems remembering that an action was taken or that an action is required	Seldom forgets own actions taken or that an action is required	Is very good at remembering own actions taken or that an action is required (e.g., change of course to avoid restricted area)

1	2	3	4	5	6	7

PRIORITIZING

Taking early or prompt action on air traffic problems rather than waiting or getting behind; knowing what to do first and identifying the most important situations; recognizing that some problems or situations are less important and can wait; preplanning before busy periods; organizing the board and using flight strips effectively to keep priorities straight for handling air traffic situations; quickly and decisively determining appropriate priorities.

Has difficulty recognizing which air traffic problems are the most pressing; may deal with problems in chronological order, or take the easy ones first	Usually recognizes the most important air traffic problems and handles them before the less pressing ones	Always recognizes which air traffic problems need immediate attention and handles them before less pressing ones; recognizes appropriate priorities for control actions
Often acts on air traffic problems without evaluating the possible consequences of these actions	Normally looks ahead to assess potential air traffic problems that might result from own actions or from changing conditions	Is very good at looking ahead to assess potential problems that might result from revised clearances, aircraft counts or altitude changes
Often puts off decisions or actions that should be taken right away	Is usually good about taking early or prompt action on air traffic problems; may sometimes put off a decision or an action that should be attended to immediately	Consistently takes early or prompt action on air traffic problems

1	2	3	4	5	6	7

MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW

Reacting to and resolving potential conflictions effectively and efficiently; using proper air traffic separation techniques effectively to ensure safety; sequencing aircraft effectively for arrival or departure; sequencing aircraft to ensure efficient/timely traffic flow; controlling traffic in a manner that ensures efficient traffic flow; controlling traffic in a manner that minimizes traffic problems (e.g., conflictions, traffic flow problems) for other controllers and pilots.

Sometimes fails to maintain minimum separation or to recognize and resolve potential conflictions	Typically uses appropriate control actions to maintain proper separation or to resolve potential conflictions	Consistently maintains safe, efficient, and orderly traffic flow, even under difficult or unusual circumstances (e.g., extremely heavy traffic)
Uses control actions that fail to resolve potential conflictions or that result in excessive workload (e.g., waits until potential conflictions are critical before taking action)	Resolves simple conflictions and traffic flow problems quickly without causing unnecessary delays	Recognizes potential problems or conditions early and takes appropriate actions to maintain separation and minimize inconvenience
Does not always sequence aircraft adequately or ensure proper spacing between aircraft; may cause excessive and unnecessary delays by choosing poor control actions, waiting too long to provide needed commands, etc.	Generally uses correct procedures to sequence and space aircraft safely; maintains smooth traffic flow, but may not use the most efficient control actions (e.g., may not always take aircraft types into account)	Sequences and spaces traffic effectively and efficiently even when extremely busy; always maintains proper separation while minimizing delays (e.g., avoids delaying vectors as appropriate, uses flow control procedures when necessary)

1	2	3	4	5	6	7

ADAPTABILITY AND FLEXIBILITY

Reacting effectively to difficult equipment problems, changes in weather, traffic situations, etc. or to unexpected actions on the part of other controllers or pilots; using contingency or fall-back strategies effectively when unforeseen/unanticipated air traffic problems emerge or if first plan doesn't work; asking for help when it's needed; developing/executing innovative solutions to air traffic problems; dealing effectively with situations for which there may not be clearly prescribed procedures or situations which require novel thinking; adapting to equipment updates, new procedures, etc.

Does not adjust well to unusual and difficult air traffic situations	Is usually able to adapt effectively to difficult situations such as rapidly worsening weather, equipment problems, etc.	Reacts very effectively to complicating events and difficult equipment problems
Rarely displays good "fall-back" strategies for dealing with unanticipated air traffic problems	Frequently, but not always, has effective contingency strategies for unforeseen or unanticipated air traffic problems when they arise	Is very adept at using effective contingency or "fall-back" strategies when unforeseen or unanticipated air traffic problems arise
Is ineffective at handling air traffic situations with no clearly prescribed procedures	For the most part, is good at handling air traffic situations that have no "textbook answers," but does better with the more routine problems	Deals very effectively with air traffic situations where there are no clearly prescribed procedures

1	2	3	4	5	6	7

Appendix C
Questionnaires

BACKGROUND QUESTIONNAIRE

Observer Code _____ Date _____

INSTRUCTIONS

This questionnaire is designed to obtain information about your background as an air traffic control specialist. The information will be used to describe the participants in this study as a group in written or oral reports. Your identity will remain anonymous, so do not write your name on the form. Instead, your data will be identified by an observer code known only to yourself and the researchers conducting this study.

---

1. What is your job position or title?

   _____

2. What is the level of your facility?

   1 2 3 4 5

3. What is your age?

   _____ years

4. How many years have you worked as an air traffic control specialist?

   _____ years

5. How many of the past 12 months have you actively controlled traffic?

   _____ months

6. How many years of experience do you have training and evaluating air traffic controllers?

   _____ years

7. Please briefly describe your air traffic control training and evaluation experience.

FINAL QUESTIONNAIRE

Observer Code _____ Date _____

A. Indicate the importance of the 6 performance areas to overall air traffic control performance by selecting a weight score (between 0 and 100) for each area. Higher weights indicate more important performance areas. Your overall performance rating for each area will be multiplied by your indicated weight to compute a weighted overall performance score for each scenario. The weights must sum to 100.

EXAMPLE:

20	MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW
20	MAINTAINING ATTENTION AND SITUATION AWARENESS
20	PRIORITIZING
20	PROVIDING CONTROL INFORMATION
10	TECHNICAL KNOWLEDGE
10	COMMUNICATING
100	

YOUR SELECTIONS:

_____	MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW
_____	MAINTAINING ATTENTION AND SITUATION AWARENESS
_____	PRIORITIZING
_____	PROVIDING CONTROL INFORMATION
_____	TECHNICAL KNOWLEDGE
_____	COMMUNICATING
100	

B. Rank order the effectiveness of the 4 controllers viewed on the videotapes by placing a 1, 2, 3, or 4 (1-highest, 4-lowest) beside each controller code number.

For ACY

C#1	C#2	C#3	C#4
_____	_____	_____	_____

For GEN

C#1	C#2	C#3	C#4
_____	_____	_____	_____

On Both Sectors

C#1	C#2	C#3	C#4
_____	_____	_____	_____

Videotape evaluations of controllers is a new methodology that has not been done in previous research. In order to evaluate and improve this methodology, we would like your opinions regarding the following questions.

1. As compared to viewing controllers "live," the videotapes showed sufficient information for me to make my evaluations.

1	2	3	4	5	6	7	8	9	10
strongly disagree									strongly agree

2. The training period was sufficient for me to become familiar with the new evaluation form.

1	2	3	4	5	6	7	8	9	10
strongly disagree									strongly agree

3. Please write down any recommendations you have for improving the videotape evaluations methodology (e.g., training format, videotape presentation, etc.).

4. Please list any other objective performance measures that should be collected to evaluate controller effectiveness (e.g., aircraft flight time, aircraft fuel consumption).

5. How can R&D help operations at your facility?

Appendix D
Correlations Between Rating Scales

Table D1. Correlations With Maintaining Safe And Efficient Traffic Flow

MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW	0.76 0.71 0.90
	0.76 0.68 0.38 0.50 0.83
	0.68 0.64 0.71 0.58 0.72
	0.44 0.43 0.54
	0.62 0.38 0.65
	0.57 0.65 0.55 0.64
	0.85

Table D2. Correlations With Maintaining Attention And Situation Awareness

	Maintaining Awareness of Aircraft Positions R5	Ensuring Positive Control R6	Detecting Pilot Deviations from Control Instructions R7	Correcting Own Errors in a Timely Manner R8	Overall Attention and Situation Awareness Scale Rating R9
MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW					
R1-Maintaining Separation and Resolving Potential Conflicts	0.76	0.68	0.38	0.50	0.83
R2-Sequencing Arrival and Departure Aircraft Efficiently	0.73	0.72	0.43	0.64	0.83
R3-Using Control Instructions Effectively	0.78	0.73	0.43	0.65	0.83
R4-Overall Safe and Efficient Traffic Flow Scale Rating	0.83	0.74	0.40	0.62	0.89
MAINTAINING ATTENTION AND SITUATION AWARENESS					
R5-Maintaining Awareness of Aircraft Positions	.	0.65	0.32	0.61	0.86
R6-Ensuring Positive Control	0.65	.	0.43	0.53	0.81
R7-Detecting Pilot Deviations from Control Instructions	0.32	0.43	.	0.45	0.62
R8-Correcting Own Errors in a Timely Manner	0.61	0.53	0.45	.	0.75
R9-Overall Attention and Situation Awareness Scale Rating	0.86	0.81	0.62	0.75	.
PRIORITIZING					
R10-Taking Actions in an Appropriate Order of Importance	0.78	0.59	0.23	0.62	0.80
R11-Preplanning Control Actions	0.69	0.59	0.41	0.68	0.75
R12-Handling Control Tasks for Several Aircraft	0.81	0.61	0.36	0.66	0.83
R13-Marking Flight Strips while Performing Other Tasks	0.65	0.50	0.27	0.60	0.67
R14-Overall Prioritizing Scale Rating	0.79	0.64	0.32	0.68	0.83
PROVIDING CONTROL INFORMATION					
R15-Providing Essential Air Traffic Control Information	0.48	0.48	0.12	0.31	0.44
R16-Providing Additional Air Traffic Control Information	0.41	0.37	0.28	0.30	0.45
R17-Overall Providing Control Information Scale Rating	0.52	0.52	0.29	0.37	0.55
TECHNICAL KNOWLEDGE					
R18-Showing Knowledge of LOAs and SOPs	0.59	0.40	0.22	0.39	0.57
R19-Showing Knowledge of Aircraft Capabilities and Limitations	0.31	0.54	0.35	0.26	0.47
R20-Overall Technical Knowledge Scale Rating	0.57	0.48	0.21	0.39	0.62
COMMUNICATING					
R21-Using Proper Phraseology	0.55	0.60	0.08	0.48	0.64
R22-Communicating Clearly and Efficiently	0.64	0.57	0.25	0.52	0.69
R23-Listening to Pilot Readbacks and Requests	0.52	0.58	0.42	0.48	0.61
R24-Overall Communicating Scale Rating	0.60	0.59	0.18	0.49	0.67
WEIGHTED OVERALL PERFORMANCE SCORE	0.86	0.81	0.47	0.71	0.93

Table D3. Correlations With Prioritizing

@PJL EOJ NAME="Microsoft Word - 9616TEXT.DOC"
@PJL EOJ NAME="NWQS038F"
34SX@PJL USTATUS JOB=ON

	R10 Taking Actions in an Appropriate Order of Importance	R11 Preplanning Control Actions	R12 Handling Control Tasks for Several Aircraft	R13 Marking Flight Strips while Performing Other Tasks	R14 Overall Prioritizing Scale Rating
MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW					
R1-Maintaining Separation and Resolving Potential Conflicts	0.68	0.64	0.71	0.58	0.72
R2-Sequencing Arrival and Departure Aircraft Efficiently	0.72	0.72	0.74	0.59	0.75
R3-Using Control Instructions Effectively	0.73	0.75	0.77	0.66	0.80
R4-Overall Safe and Efficient Traffic Flow Scale Rating	0.81	0.76	0.84	0.67	0.84
MAINTAINING ATTENTION AND SITUATION AWARENESS					
R5-Maintaining Awareness of Aircraft Positions	0.78	0.69	0.81	0.65	0.79
R6-Ensuring Positive Control	0.59	0.59	0.61	0.50	0.64
R7-Detecting Pilot Deviations from Control Instructions	0.23	0.41	0.36	0.27	0.32
R8-Correcting Own Errors in a Timely Manner	0.62	0.68	0.66	0.60	0.68
R9-Overall Attention and Situation Awareness Scale Rating	0.80	0.75	0.83	0.67	0.83
PRIORITIZING					
R10-Taking Actions in an Appropriate Order of Importance	.	0.78	0.88	0.77	0.92
R11-Preplanning Control Actions	0.78	.	0.82	0.79	0.91
R12-Handling Control Tasks for Several Aircraft	0.88	0.82	.	0.78	0.93
R13-Marking Flight Strips while Performing Other Tasks	0.77	0.79	0.78	.	0.87
R14-Overall Prioritizing Scale Rating	0.92	0.91	0.93	0.87	.
PROVIDING CONTROL INFORMATION					
R15-Providing Essential Air Traffic Control Information	0.39	0.52	0.48	0.42	0.48
R16-Providing Additional Air Traffic Control Information	0.32	0.47	0.42	0.36	0.43
R17-Overall Providing Control Information Scale Rating	0.49	0.60	0.55	0.52	0.57
TECHNICAL KNOWLEDGE					
R18-Showing Knowledge of LOAs and SOPs	0.56	0.48	0.56	0.48	0.56
R19-Showing Knowledge of Aircraft Capabilities and Limitations	0.30	0.36	0.28	0.18	0.32
R20-Overall Technical Knowledge Scale Rating	0.56	0.48	0.55	0.41	0.57
COMMUNICATING					
R21-Using Proper Phraseology	0.64	0.60	0.61	0.58	0.68
R22-Communicating Clearly and Efficiently	0.73	0.73	0.75	0.77	0.78
R23-Listening to Pilot Readbacks and Requests	0.64	0.68	0.64	0.67	0.69
R24-Overall Communicating Scale Rating	0.70	0.67	0.70	0.72	0.75
WEIGHTED OVERALL PERFORMANCE SCORE	0.83	0.83	0.86	0.76	0.89